PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/50

A2

(11) International Publication Number: WO 00/60507

(43) International Publication Date: 12 October 2000 (12.10.00)

(21) International Application Number: PCT/US00/08777

(22) International Filing Date: 31 March 2000 (31.03.00)

(30) Priority Data:
60/127,486 2 April 1999 (02.04.99) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 60/127,486 (CIP)
Filed on 2 April 1999 (02.04.99)

(71) Applicant *(for all designated States except US)*: NEOGENESIS, INC. [US/US]; 840 Memorial Drive, Cambridge, MA 02139 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: WINTNER, Edward, A. [US/US]; 44 Valentine Street, Cambridge, MA 02139 (US). MOALLEMI, Ciamac, C. [US/US]; Apartment 2–4A, 100 Memorial Drive, Cambridge, MA 02142 (US).

(74) Agent: KLUNDER, Janice, M.; Hale and Dorr, LLP, 60 State Street, Boston, MA 02109 (US).

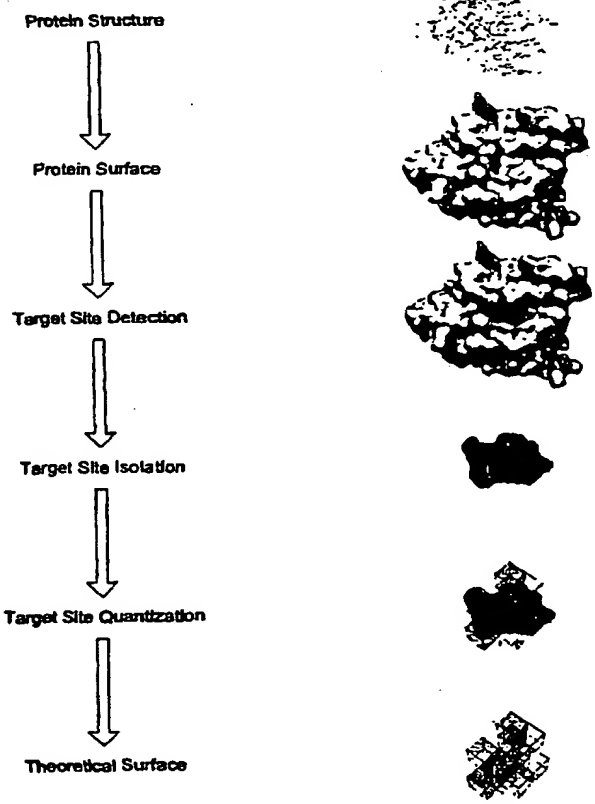
(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: ANALYZING MOLECULE AND PROTEIN DIVERSITY

(57) Abstract

A computer–based method in which a set of constraints is placed on possible target surfaces, and a fully enumerated set of theoretical target surfaces under the given constraints is created, such that each surface has a defined, continuous volume and a defined, continuous surface area. One or more sets of objects are mapped to the fully enumerated set of theoretical target surfaces to define corresponding subsets of the fully enumerated set of theoretical target surfaces. An aspect of diversity of the objects is analyzed based on degrees of similarities and differences among the corresponding subsets.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

ANALYZING MOLECULE AND PROTEIN DIVERSITY

This application claims priority from Provisional United States Patent Application Serial Number 60/127,486, filed April 2, 1999, and incorporated by reference.

This invention relates to analyzing molecule and protein diversity.

Combinatorial chemistry allows the creation of unprecedented numbers of organic compounds. The rational synthesis of millions of small organic molecules is now achievable in a matter of days. There are estimated to be more than ten to the hundredth power of small molecules that could be synthesized using current methods. By "small", we mean a molecule having fewer than 1500 Daltons, where a Dalton is defined as 1/12 of the weight of a carbon 12 atom or roughly the weight of a hydrogen atom.

An important question is how can one create a set of molecules of such diversity as to contain at least one potent binder to any given target of interest? This question is central to drug discovery in an era that is characterized by a growing wealth of DNA sequence information and a relative dearth of corresponding target structures and their functions. In a world in which there are many more putative targets than can be studied by x-ray crystallography, multi-dimensional NMR, or other high resolution biophysical techniques, any attempt to generate biologically active ligands to targets of unknown structure will require general screening libraries: libraries of molecules that cover a high percentage of so-called "diversity space".

The utility of small molecules as drugs depends in part on molecular complementarity: how well the molecules fit and/or stick to chemically active sites (often in the form of depressions in a protein) on the surface of a cell, on the surface of an intracellular organelle, or on a cytosolic protein. The potential molecular complementarity of a small molecule is in large part determined by two factors:

1) The shape of the molecule, meaning the total Van Der Waals (VDW) surface of a given conformation of the molecule and how it follows or does not

follow the VDW surface of the target site of interest. The shape complementarity of molecule and target are largely responsible for energetic forces such as the displacement of water (the "hydrophobic effect") and so called "Van Der Waals" or "London dispersion" forces.

2) The types of potential energetic interactions (such as hydrogen bonding, percentage ionic bonding, proximity of polarizable moieties) found at various places on the molecule, and on the manner, order, and spatial orientation in which the energetically interactive portions of the molecule are connected to each other and presented in the presence of the target of interest.

Typically, each of these factors is measured or calculated in only a relative way with respect to a specific set of molecules or with respect to specific protein surfaces being examined. A general comparison of potential molecular complementarity between two sets of molecules, however, requires doing calculations or experiments using an absolute or fixed frame of reference.

For example, if company 1 has tested a molecule set A for affinity to a set of target surfaces X found in cancerous cells, and company 2 has tested a molecule set B against a set of target surfaces Y found in nervous tissue, the value of molecule set B with respect to the target surfaces X of company 1 is not apparent because each of the affinity evaluations has been performed against a different standard. Also, without a full molecular calculation of set A against target surfaces Y, it is not apparent whether potential chemically active portions of the target surface Y could be bonded by molecules in set A. Finally, even if all calculations of sets A and B vs. surfaces X are performed, neither company will gain a measure of the "absolute diversity" of their molecule sets; that is, they will have no measure of the likelihood that either sets A or B will contain a molecule that has potential activity against any given target T. This is because the standard against which they have measured their molecules are only a small subset of potential target surfaces.

SUMMARY OF THE INVENTION

Implementations of the invention provide an absolute or fixed frame of reference for comparing sets of molecules and protein surfaces. By defining molecules in terms of their complementarity or attraction to a fully enumerated basis set of theoretical protein surfaces within given parameters, it is possible to measure the diversity of a set of molecules against a non relative standard, i.e., an absolute measurement. This allows for an efficient comparison of different sets of molecules, for example, sets of drugs, and enables meaningful categorization of classes of molecules against a standard set of surfaces. Furthermore, this allows for the detection of theoretical protein surfaces to which no molecules in a set are complementary, thus enhancing the ability of a chemist to design novel molecules that supplement the deficiencies of the original set. By defining real world protein surfaces in terms of their similarity to a basis set of theoretical protein surfaces, it is similarly possible to categorize real world sets of protein surfaces against a standard set of theoretical surfaces. This allows for improved classification of proteins into similar target classes by the similarity of their surface sites. By evaluating an actual set of molecules against theoretical protein surfaces and further by evaluating a set of real world protein surfaces of interest against the same theoretical protein surfaces, it is possible to select classes of molecules that are likely to have substantial activity against the real world protein surfaces. This allows for improved molecule screening, for example, in drug research. By evaluating an actual set of molecules against theoretical protein surfaces and further by evaluating a set of real world protein surfaces of interest against the same theoretical protein surfaces, it is also possible to find actual protein surfaces to which no molecule in the set A has likely activity. This enhances the ability of a chemist to design molecules beyond those in set A that match the previously unmatched protein surfaces of interest, and may thereby be tested for pharmaceutical activity, thus supplementing deficiencies in the original set of molecules.

Thus, in general, in one aspect, the invention features a computer-based method in which a set of constraints on possible target surfaces is defined, and a fully enumerated set of theoretical target surfaces under the defined constraints is

also defined, such that each surface has a defined, continuous volume and a defined, continuous surface area. One or more sets of objects are mapped to the fully enumerated set of theoretical target surfaces to define corresponding subsets of the fully enumerated set of theoretical target surfaces. An aspect of diversity of the objects is analyzed based on degrees of similarities and differences among the corresponding subsets.

Implementations of the invention may include one or more of the following features. The target surfaces may include negative space target surfaces. The objects may include positive space object surfaces associated with different molecules. The objects may be mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which positive space object surfaces of conformations of molecules are complementary. The aspect of diversity that is analyzed may be the difference or similarity between the molecules which map to those negative space theoretical target surfaces.

The objects may include negative space object surfaces associated with different proteins, and the objects may be mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which negative space object surfaces of protein pockets are similar. The aspect of diversity that is analyzed may be the difference or similarity between protein pockets which map to those negative space theoretical target surfaces. The objects may include positive space object surfaces associated with different molecules and negative space object surfaces associated with different proteins. In the case of molecules, the objects may be mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which positive space object surfaces of conformations of molecules are complementary. In the case of proteins, the objects may be mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which negative space object surfaces of protein pockets are similar. The aspect of diversity that is analyzed may be the difference or similarity of the molecules which map to those negative space theoretical target

surfaces to the protein pockets which map to those negative space theoretical target surfaces.

The theoretical target surfaces and the objects may be polyhedrons, e.g., cubes, all of the same size and shape. The set of all theoretical target surfaces defines a diversity space within which the diversity of objects can be measured by mapping those objects to the diversity space. Regions of the diversity space to which no objects map may be identified, and molecules may be designed that occupy at least one of the unfilled theoretical target surfaces of the diversity space.

Complementarity may be associated with binding affinities of positive space object surfaces of conformations of molecules to negative space theoretical target surfaces.

The constraints may include volume, associations of each of a number of sites of the target surface with a preselected molecular property drawn from a larger set of possible molecular properties, including hydrophobic, polarizable, H-bond acceptor, H-bond donor, H-bond donor/acceptor, potentially positively charged, and potentially negatively charged. Fewer than all of the sites of the target surface may each be associated with a different one of the molecular properties and all of the other sites of the target surface may be associated with a common molecular property, such as slightly hydrophobic. The degrees of similarities or differences may involve functional properties associated with the corresponding subsets of the fully enumerated set of theoretical target surfaces or shape properties associated with the corresponding subsets of the fully enumerated set of theoretical target surfaces.

Each of the objects may be defined by quantizing molecules into polyhedrons. Each of a fixed set of orientations of each conformation of each of the objects may be fitted to each of the target surfaces, and each of the fittings may be scored.

The constraints may include a resolution of the polyhedrons, e.g., 4.24 Angstroms, or maximum and minimum numbers of polyhedrons.

Each of the polyhedrons may share a common interface with another of the polyhedrons. The constraints may also include the absence of any occluded volumes greater than a given user-defined parameter. The target surfaces may be defined conceptually as having been carved out of a flat surface.

In general, in another aspect, the invention features categorizing existing molecules based on negative space target surfaces to which conformations of the molecules are complementary, and designing novel molecules that are complementary to negative space target surfaces to which no conformations of the existing molecular are complementary.

In general, in another aspect, the invention features a method of creating novel molecules to be tested as ligands for proteins. In the method, proteins are categorized based on target surfaces to which their pockets of known structure map, and novel molecules are designed that are complementary to the negative space target surfaces to which the protein pockets map.

In general, in another aspect, the invention features a computer programmed to determine the chemical similarity of different molecules. The program approximates the surface shape of each one of a plurality of molecules of interest by linking a series of cubes, each cube having a dimension R, the locations of the cubes being determined by the calculated electron probability density of the individual one of the molecules of interest, each cube sharing at least one of its six faces with another cube, such that there is a specific number of linked cubes which varies for each individual one of the plurality of molecules of interest. The chemical reactivity of each individual one of the plurality of molecules of interest is approximated by assigning each cube of each individual one of the plurality of molecules of interest, no more than one functionality value from a plurality of M different chemical functionality values. The surface shape and chemical reactivity of a chemically active surface having a volume equal to V is approximated by subtracting a number V/R^3 cubes of dimension R from a surface, wherein each of the cube spaces shares at least one face with another cube space and wherein N of the cube spaces has one of a plurality of M different chemical functionality values. An attraction value K is calculated for each one of

the plurality of molecules of interest to the chemically active surface. A list of overall attraction values to the chemically active surface is calculated.

Implementations of the invention may include one or more of the following features. The calculation of the attraction value K may be performed on a plurality of different predetermined chemically active surfaces, and a matrix of overall attractive values of each molecule of interest to each of the different surfaces may be calculated. The molecules of interest may include organic molecules. The chemically active surface having a plurality of predetermined active chemical locations may be calculated to correspond to the shape of an actual protein surface structure. The molecules of interest may be organic molecules of 1500 Daltons or less. The chemically active surface having a plurality of predetermined active chemical locations may be compared to an actual protein surface to calculate a similarity value of the actual protein surface to the predetermined active chemical locations. The predetermined chemically active surfaces may be compared to a plurality of actual protein surfaces and a matrix of similarity values may be calculated. The cube spaces subtracted from the surface may be calculated to approximate the electron probability density of at least one of a plurality of depressions in known protein surface structures. The N sites of chemical functionality may be calculated to approximate the location and type of chemical functionality of actual depressions in known protein structures.

Other advantages and features will become apparent from the following description and from the claims.

DESCRIPTION

Figure 1 shows a $(CH_2)_n$ chain encapsulated by 4.24 A cubic units.

Figure 2 shows examples of surfaces allowed and disallowed by the non-occlusion parameter in a theoretical target surface generation algorithm. Gray shading represents the opening of the theoretical surface. A is allowed. B is disallowed due to two occluded negative space cubes (marked X).

Figure 3 shows a theoretical target surface of 13 negative space cubes and four sites of specific molecular property interaction: hydrophobic (white), polarizable (purple), H-bond accepting (green), and H-bond donating (orange). Blue shading indicates the opening of the theoretical surface.

Figure 4 shows a "quantized" representation (Q-file) of one conformation of molecule 6a superimposed on its atomic structure (ball and stick and space-filling model). Molecular property characteristics of the Q-file are hydrophobic (white quanta), polarizable (purple quanta), H-bond accepting (green quanta), and negatively charged (red quanta).

Figure 5 shows test molecules.

Figure 6 shows a ranking of molecules by QCSD similarity scores.

Figure 7 illustrates examination of a theoretical target surface common to molecules 8c (top) and 8a (bottom). Blue shading indicates opening of the theoretical surface. Specific points of complementarity on the theoretical target surface are hydrophobic (white) polarizable (purple) and H-bond donating (orange). Superimposition of the original molecular conformations onto the theoretical target surface demonstrates that the extra phenyl substituent of 8c protrudes from the opening of the theoretical surface and is not involved in complementarity to the surface.

Figure 8 illustrates examination of a theoretical target surface common to molecules 1a (A) and 5a (B). Blue shading indicates opening of the theoretical surface. Specific points of surface complementarity found by QSCD are hydrophobic (white), polarizable (purple), and H-bond donating/accepting (yellow). Overlay plot (C) of the non-hydrogen backbones of 1a (orange) and 5a (green) indicate similar features. Aromatic/hydrophobic overlaps are shown in purple; H-Bond donating oxygens are in red. Overlay generated with Sybyl version 6.5 (Tripos Inc, 1699 S. Hanley Rd., St. Louis, MO, 63144).

Figure 9 illustrates ranking of molecules in Figure 5 by Tanimoto similarity score of 2D UNITY fingerprints.

Figure 10 shows a QSCD plot of all of the theoretical surface shapes covered by all of the conformations of all of the molecules (blue dots) in Figure

5. The total volume of the cube encompasses all 49,268,918 theoretical surface shapes as listed in Table 1. Red dots show two exemplary theoretical surface shapes (a, b) not covered by any of the molecules in Fig. 5. Axes used are functions of opening area, opening length/width, and depth per opening quantum.

Figure 11 shows a map of the 20 compounds in Fig. 5 (blue dots) in a representative BCUT three-axis diversity space. BCUT axes used are, respectively: 1) BCUT HACCEPT S INVDIST 050 R H, 2) BCUT HDONOR S INVDIST 030 R H, and 3) BCUT TAB POLAR S INVDIST 300 R L. Red dot shows an unfilled coordinate of diversity space, at (7.54, 7.25, 6.82). The information contained in this BCUT coordinate does not reveal information about the shape of a molecule which might be able to fill this position in diversity space.

Figure 12 shows use of QSCD to design complementary combinatorial libraries to unmatched theoretical target surfaces. Many conceivable libraries of a given shape and functionality may be designed to fill a given unmet diversity need.

Figure 13 shows two sample surfaces.

Figure 14 illustrates the quantization process.

Figure 15 shows a legend for symbols used in functionality rule diagrams.

Figure 16 shows functionality rules for Potential Negative Charge Functionality. The structures are searched for in order.

Figure 17 shows functionality rules for Potential Positive Charge Functionality. The structures are searched for in order.

Figure 18 shows a functionality rule for Hydrogen Bond Donor/Acceptor Functionality.

Figure 19 shows a functionality rule for Hydrogen Bond Donor Functionality.

Figure 20 shows a functionality rule for Hydrogen Bond Acceptor Functionality. The structures are searched for in order.

Figure 21 shows a functionality rule for Polarizable Functionality.

Figure 22 shows Table 5, a ranking of molecules in Fig. 5 by QSCD diversity score. Blue = homogeneous pairs, yellow = +phenyl pairs (8c), green = ATI-AT2 pairs (3,4)

Figure 23 shows Table 6, a ranking of molecules in Fig. 5 by Tanimoto similarity score of 2D UNITY fingerprints. Blue = homogeneous pairs, yellow = +phenyl pairs (8c), green = ATI-AT2 pairs (3,4)

Figure 24 shows an example subset of theoretical surfaces Ti containing 4 members and an example central set Ci for Ti (F = 1, E = 3) where the black face denotes a point of attachment A1 on Ci:

Figure 25 shows an example Core Molecule Mi to fill Central set Ci.

Figure 26 shows an example Library (Mi,B) where B = a set of amines.

Figure 27 shows an example subset of target surfaces Ti containing 4 members and an Example Central set Ci for Ti (F = 1, E = 3), where the black face denotes a point of attachment A1 on Ci:

Figure 28 shows an Example Core Molecule Mi to fill Central set Ci.

Figure 29 shows an example Library L(Mi, B) where B = a set of amines.

Figure 30 shows a protein quantization process.

We define diversity as the measure, based on pre-defined criteria, of the difference or similarity among all members of a set. In a pharmaceutical setting, molecular diversity can be defined as the measure, based on biological criteria, of the difference or similarity between small molecules. Each of the existing methods of calculating biologically relevant diversity of small molecules defines slightly different criteria for molecular comparison, and thus a different configuration of diversity space as a whole. Examples include low dimensional diversity space such as BCUT metrics, high dimensional diversity space such as Chem-X/ChemDiverse multiple point pharmacophores, and empirical biological diversity space such as affinity fingerprinting.

Many known ways of quantifying the diversity of molecules use molecular properties such as functionality and connectivity as a basis for

categorization (see for instance Potter and Matter, *J. Med. Chem.*, 1998, p. 478). For example, in the BCUT method used to generate four- to six-dimensional diversity space, molecules are broken down into matrices according to connectivity and molecular interaction properties. Coordinates in diversity space are assigned through the resulting eigenvalues of these matrices, leading to useful multi-dimensional plots of molecular diversity. However, because the use of eigenvalues is an irreversible transformation (different 3D shapes can map to the same eigenvalues), it follows that an empty coordinate in BCUT diversity space cannot be translated into a 3D template of a "missing molecule." Thus, while a model such as BCUT diversity is well validated as a tool for finding combinatorial matches to a lead compound or pharmacophore, it cannot be directly used to populate the entire diversity space that it defines.

Similarly, in the popular Chem-X/ChemDiverse diversity package, molecules are broken down into all accessible three- or four-point pharmacophores of triangular or tetrahedral functionality distances. If the model is used to display molecular diversity, coordinates in diversity space are assigned through the resulting string of accessible three- or four-point pharmacophores; this method has been shown to be highly effective in classifying molecules by pharmacological similarity. However, the mapping of complex 3D shapes to a set of triangular or tetrahedral functionality distances is an irreversible transformation; empty three- or four-point pharmacophores in Chem-X derived diversity space cannot be translated into a 3D template of a complex shape. Since a set of coordinates in Chem-X is insufficient to define the shape of "missing molecules," Chem-X cannot be used to directly populate empty molecular diversity space.

Another example of current diversity methods is affinity fingerprinting, in which molecules are empirically assayed against a panel of 10-20 actual proteins selected to be promiscuous in their ability to bind small molecules. Position in molecular diversity space is assigned through the resulting string of IC_{50} binding values, and these affinity fingerprints provide unprecedented ability to group similarly active compounds in diversity space. However, because the actual mode

of binding in any assay is not incorporated in the resulting IC_{50} value, the mapping of molecules to the selected protein panel is an irreversible transformation. Thus, an empty coordinate in affinity fingerprinting diversity space (an "unmatched" string of IC_{50}s to a given protein panel) cannot be back-translated into a 3D molecular template. A similar affinity fingerprinting diversity method has been put into practice using a panel of computational surfaces of real-world protein pockets and a modified form of the DOCK program. While this method shows similar promise in its ability to detect pharmacological similarity, it is, like its empirical affinity fingerprinting counterpart, an irreversible mapping.

Thus, for the most part, current methods are able to successfully identify compounds of the same pharmacological class as being similar and compounds of different pharmacological classes as being different. Given a starting pharmacophore from known ligands and/or the target site of a target crystal structure, such methods interface well with the design of complementary combinatorial libraries.

The design of combinatorial libraries to cover all of diversity space is a rather different problem, however. In this case, it is not enough to be able to compare existing molecules for differences or similarities. In addition to being able to place molecules relative to one another in diversity space, one must be able to point to an absolute area of diversity space not yet covered and from its coordinates design a novel set of compounds to fill that uncovered space.

In order to rationally and systematically fill diversity space, an informationally reversible diversity model is needed. This model must be formulated such that: members (in this case molecules) can be assigned to coordinates for similarity/dissimilarity comparison, and empty coordinates retain the information necessary to directly generate coordinate membership.

One good path for such a model is to use as coordinates the exact information that differentiates one member from another, without intervening, irreversible transformations. To apply this reasoning to molecular diversity, it must first be asked: what are the criteria by which diversity of compounds is to be measured (what information differentiates one molecule from another). One of

the most fundamental criterion in molecular drug discovery is the extent to which two molecules have similar or different binding affinities to a given target. With the assumption that similar binding affinity tracks with a molecule's complementarity to similar target surfaces, we have selected as our criterion for diversity complementarity to a fully enumerated set of theoretical target surfaces.

Given the above definition of molecular diversity, it remains to provide parameters under which to define a biologically relevant basis set of enumerated theoretical target surfaces and to quantify molecular complementarity to a given theoretical target surface at a level which is both in accordance with known principles of molecular recognition and computationally applicable to millions of compounds. With a numerical determination of complementarity and a biologically relevant basis set of surfaces, molecular diversity space is thus absolutely established as the molecular complement to a fully enumerated set of theoretical target surfaces.

We introduce the concept quantized surface complementarity diversity (QSCD), which defines a molecule numerically by a mapping that describes its complementarity K to every distinct theoretical protein surface of resolution R not exceeding volume V with N sites of M types of chemical functionality P_{MN}. K is defined as an algorithm that takes into account the molecular shape and chemical functionality of both the given molecule and the given theoretical protein surface. From this definition, it follows that a comparison of two molecules will yield a numerical difference that is representative of their complementarities: to the extent that two molecules each have complementarity for the same theoretical protein surfaces, the molecules are similar; to the extent that two molecules have no complementarity to common theoretical protein surfaces, the molecules are dissimilar. In other words, "similarity" between molecules is defined as the ability to complement the same theoretical protein surfaces and "difference" between molecules is defined as the ability to complement different theoretical protein surfaces.

Because QSCD uses complementarity to theoretical protein surfaces as a basis for categorization, both 3-D shape and molecular functionality are taken

into account. Also, because QSCD uses as its basis a complete set of theoretical protein surfaces (under R, V, and P_{MN}), the method provides diversity information in a fixed frame of reference: coordinates in quantized surface complementarity diversity space are independent of any molecules or natural protein surfaces compared in that space. Because of this, not only can the diversity of disparate sets of molecules be compared without having to compare the molecules to each other directly, but the diversity of any set of actual natural protein surfaces can be examined through complementarity to the theoretical basis set. In addition, complementarity of molecules to a set of theoretical protein surfaces representative of actual natural protein surfaces can be examined in the context both of any other set of molecules and any other set of protein surfaces. Furthermore, given a set of molecules, it becomes immediately apparent what percentage of theoretical protein surfaces are covered by complementary molecules, thus giving a measure of the set's molecular diversity in the space defined by all potential surfaces under R, V, and P_{MN}. Not only does this provide a measure of diversity in an absolute sense which is not relative to any historically biased set of surfaces or molecules, but it also makes clear a set of theoretical target surfaces to which no molecules in the initial set bind, allowing a straightforward design of novel molecules to supplement the initial set.

Theoretical Target Surfaces

In one implementation, to generate a finite set of theoretical target surfaces that approximates all possible binding pockets with volume equal to or less than V, we consider each theoretical surface to be formed by successively carving c ubic units out of an initially flat surface. These cubic units represent "negative space" that a potential ligand could occupy. Given cubic units with sides of length R (the resolution of the model), we use at most V/R^3 negative space cubes to describe each theoretical target surface. Others have previously employed cubic units to successfully approximate complementarity between small molecules and individual protein surfaces.

The size of a negative space cube is directly related to the resolution and type of diversity data which the user desires as output. In choosing the size of the

negative space cube, one motivation is to maximize negative space cube size such that the difference of a single cube in a surface is highly differentiating in terms of molecular recognition (i.e., every surface is orthogonal to every other surface). At the same time, enough information must be retained in each negative space cube to predict shape and functional complementarity at a ligand/surface interface. The former constraint minimizes overlap of diversity information while the latter constraint maximizes precision of diversity information. Together, the competing constraints result in a basis unit for the enumeration of theoretical target surfaces that minimizes the number of negative space cubes needed to accurately model diversity for a given volume V.

A resolution of 4.24 Å negative space cubes was found by computer optimization of test molecules to provide an upper limit of cube size while still maintaining an acceptable level of molecular shape information. Interestingly, 4.24 Å is the approximate VDW "cross-section" of a (CH2)n chain; a series of 4.24 Å units neatly encapsulates a (CH2)n chain in its ground state conformation as shown in Figure 1.

In one implementation the basis set for diversity can be a set of theoretical target surfaces comprised of all possible shape combinations of 6 to 14 negative space cubes of resolution 4.24 Å (negative volume between 460 and 1070 cubic Å) subject to the following rules: Surfaces are created by successively "carving out" negative space cubes from a flat block of infinite width and depth (the theoretical target). All negative space cubes of a given surface must share at least one face with another negative space cube of the surface, and all must be part of a single, contiguous negative surface. No negative space cubes may be occluded in the +Z axis of the infinite surface block; that is, there may be no solid surface between any negative space cube and the surface plane of the infinite block. As shown in figure 2, the surface A is allowed, but the surface B is disallowed. Surfaces duplicating a previous surface with respect to rotation in the X-Y plane are discarded.

The occlusion rule provides a compromise between complete coverage of topological possibilities and acceptable computational speed. This compromise

was made based on the topological assumption that occlusions of 4.24 Å or more are infrequent in small molecule/target interactions, and that their omission would thus have only a small effect on predicting diversity of binding affinities of small molecules.

Applying the rules yields 49,268,918 unique negative surface shapes including chiral opposites. Covering a negative volume between 460 and 1070 cubic Å, these surface shapes are deemed sufficient to examine diversity of most small molecules. For instance, examining a previously published reference set of pharmaceutically relevant compounds (a filtered Comprehensive Medicinal Chemistry or CMC database), 5049 out of 5120 compounds (98.6%) have a volume of 1070 cubic Å or less.

Within each of the 49,268,918 unique negative surface shapes, each negative space cube is assigned a molecular property characteristic Pm that represents the dominant molecular environment which any atoms that are placed within that negative space will experience. Properties used are P1 hydrophobic, P2 polarizable (includes aromatics), P3 H-bond acceptor, P4 H-bond donor, P5 H-bond donor/acceptor, P6 potentially positively charged (basic), and P7 potentially negatively charged (acidic). These seven types of molecular environments are assumed to represent a minimal basis set of factors that contributes to the electrostatic/VDW complementarity of a ligand and a target surface. In one implementation, four positions of particular molecular property P1-7 are assigned, leading to 74*N!/((N-4)!*4!) surfaces for each surface shape of N negative space cubes. All other (N-4) cubes not assigned a particular molecular property are given property P8, slightly hydrophobic. The latter assignment is based on an assumption that hydrophobic effects are, on average, the largest single component contributing to ligand/target interaction.

In sum, the above process implies as a basis set for molecular diversity $1.1 * 10^{14}$ theoretical target surfaces of negative volume between 460 and 1070 cubic Å and having four sites of specific molecular property characteristics P1-7. The numerical breakdown of these 110 trillion surfaces is listed in Table 1.

Table 1: Numerical breakdown of the total number of theoretical target surfaces created using the algorithm given in the text. Surfaces consist of 6-14 negative space cubes and 4 sites of 7 possible molecular property characteristics. Number of functionally different surfaces per surface shape varies for infrequent cases in which a given shape has an axis of symmetry, so actual number of unique surfaces is slightly less than (# surface shapes) * 7^4*N!/((N-4)!*4!).

Volume (number (N) of negative space cubes)	Number of unique surface shapes	Approx. number of functionally different surfaces per unique surface shape: 7^4*N!/((N-4)!*4!)	Exact number of unique surfaces
6	212	36,015	7,163,338
7	885	84,035	73,271,443
8	3,959	168,070	655,324,488
9	17,747	302,526	5,350,917,208
10	81,407	504,210	40,912,578,322
11	375,897	792,330	297,622,676,624
12	1,753,218	1,188,495	2,082,225,979,379
13	8,224,443	1,716,715	14,116,888,070,845
14	38,811,150	2,403,401	93,264,917,290,356
Total 6-14	49,268,918	--	109,808,653,272,003

One such surface is shown in Fig. 3.

A pseudocode description of an algorithm for determining the set of theoretical target surfaces is set forth in Appendix A.

Molecular Quantization

To measure complementarity of small molecules to the basis set of theoretical target surfaces, the small molecules must be formatted in a similar frame of reference, for instance by quantizing them into positive space cubes ("quanta") of resolution 4.24 Å according to the following process (illustrated in Fig. 4):

A set of up to 100 minimized energy conformations within user-defined parameters is created. In one implementation, Tripos Multisearch modeling is used, and all conformations within 10 kcal of the lowest energy conformation found are accepted.

For each conformation, a 4.24 Å 3D grid of cubes (quanta) is aligned on top of the 3D structure using the molecule's principle axes of rotation (calculated with all atoms having mass 1).

To all 4.24 Å quanta which contain at least a user-defined % of the VDW radius of any atom, a dominant molecular property characteristic is assigned based on connectivity rules (e.g. R-[C=O]-O-H yields P_7, R-O-H yields P_5; see definitions of P_1 - P_7 above). Order of dominance is from P_7 to P_1, in order of maximum complementarity score obtainable by a given characteristic as shown in Table 2:

Table 2: Relative magnitudes of parameters used in calculating molecular property interactions between negative space cubes (theoretical target surfaces) and positive space cubes (quantized molecules). Magnitudes (listed from highest to lowest): +++, ++, +, 0, -, --, ---.

	Quantized Molecule Properties						
Theoretical Target Surface Properties	(P7) neg	(P6) pos	(P5) hb don/acc	(P4) hb donor	(P3) hb acceptor	(P2) polarizable	(P1) hydrophobic
(P7) neg charged	---	+++	0	+	-	--	--
(P6) pos charged	+++	---	0	-	+	0	--
(P5) hb don/acc	0	0	++	+	+	-	--
(P4) hb donor	+	-	+	--	++	-	--
(P3) hb acceptor	-	+	+	++	--	-	--
(P2) polarizable	--	0	-	-	-	++	0
(P1) hydrophobic	--	--	--	--	--	0	+
(P8) (surface only)	-	-	0	-	-	0	0

Minimum % of VDW radius parameter allows for a user-defined protrusion beyond the surface of a quantum cube, adding a measure of topological "flexibility" to the quantization process. A user defined 32% was found to be especially good.

The total number of 4.24 Å quanta that have been assigned a property characteristic is counted.

The grid alignment is shifted per user-defined parameters and the process is repeated until all shift combinations have been searched.

For each conformation in the original set, a "Q-file" (3D configuration of property-assigned quanta) is saved that has the lowest number of quanta in and is closest to the principle alignment.

Thus, an average molecule in this implementation is represented by 100 Q-files, each file consisting of N positive space cubes or quanta of 4.24 Å resolution having an assigned molecular property characteristic P_m (m=1-7). A typical Q-file (molecule **6a**) is shown in Fig. 4, superimposed upon its corresponding conformation. The process of optimization of quantization parameters is described later.

A pseudocode description of an algorithm for performing the quantization is set forth in Appendix B.

Mapping

Given molecules which have been rendered into sets of Q-files, each quantized conformation can be mapped into the diversity space defined by the set of 1.1×10^{14} theoretical target surfaces. In general, the following process is used.

For each quantized conformation of each molecule, each of its 24 possible X/Y/Z rotations (6 faces * 4 rotations per face) is fit to each of the 49,268,918 available surface shapes. For a given conformation-to-surface shape fit, if at least a user-defined minimum number of negative and positive space cubes overlap (in one implementation either 9 quanta or N-2 quanta of a conformation of N quanta), and if no quanta of the conformation extend beyond the bounds of the surface shape except at the mouth of the surface shape, then the complementarity of the quantized conformation to all theoretical target surfaces of that shape is examined in detail as explained next. If the above conditions are not met, the next conformation is examined.

A score is generated for the complementarity of the given conformation to each theoretical target surface of a given shape from based on user-defined parameters. (The process of optimization of the complementarity parameters is described later.) The following complementarity parameters can be used:

a) A negative parameter for each rotatable bond of the conformation.

b) If conformational energies are calculated, a negative parameter for the energy of the conformation above the lowest energy conformation from that molecule.

c) A positive parameter for the hydrophobic energy gained by removing "water" from any hydrophobic (P_1) or polarizable (P_2) surface face of either the conformation or the theoretical surface.

d) A positive parameter for the hydrophobic energy gained by removing "water" from any mildly hydrophobic (P_8) surface face of the theoretical surface.

e) A positive or negative molecular property interaction parameter for overlapping negative and positive space cubes as depicted in Table 2.

If and only if the resulting score meets a user-defined minimum, then the conformation (and thus the molecule it represents) is said to be complementary to the given theoretical target surface.

The computational advantage inherent in the process of molecule and surface quantization is realized in the speed of complementarity checking. Whereas a traditional docking program must search a high-dimensional configuration space, the implementations of the invention resolve the problem to a framework bounded by 24 possible fitting orientations and a finite number of translations. This approximation allows three-dimensional diversity computation on a scale that is applicable to very large sets of molecules.

A pseudocode description of an algorithm for performing the mapping is set forth in Appendix C.

The above process results in a complementarity map that consists of a list of all theoretical target surfaces to which at least one conformation of a molecule is complementary. Comparison of these maps provides a novel method for measuring diversity of small molecules. We term the model on which this process is based quantized surface complementarity diversity (QSCD) because it calculates diversity by measuring complementarity to a quantized representation of theoretical target surfaces.

To maintain a computationally efficient complementarity scoring system, QSCD makes many approximations of molecular recognition. As explained, these include cubic units of 4.24 Å resolution, gross approximations of surface contact area, exactly 4 points of 7 finite types of molecular property

characteristics, static theoretical surfaces, and a limited set (up to 100) of low energy conformers. Thus, the final complementarity scores are not presumed to give precise binding energies for any individual match of conformation to target surface. However, taken over all conformations of a molecule and across an enumerated set of theoretical target surfaces, the scoring system is statistically relevant as explained below.

Model Validation

To test the validity of the QCSD model, i.e., its ability to predict the extent to which two molecules have similar or different binding affinities, eight sets of test molecules were analyzed (Fig. 5), seven of which were known to have binding affinities to seven distinct targets (in addition to a known overlap between sets **3** and **4**). An eighth set with no known binding affinities was chosen with minor atomic and spatial changes to examine the sensitivity of the QCSD model at 4.24 Å resolution. Known activities of the molecules in Fig. 5 are listed in Table 3 with references.

Table 3: Pharamacological activities of the molecules used in this study (see Fig. 5). a) Doherty et al. *J. Med. Chem.* 1995, *38,* 1259-1263. b) Uehling et al. *J. Med. Chem.* 1995, *38*, 1106-1118. c) Chang et al. *J. Med. Chem.* 1994, *37*, 4464-4478. d) Chang et al. *J. Med. Chem.* 1993, *36*, 2558-2568. e) Tsutsumi et al. *J. Med. Chem.* 1994 *37*, 3492-3502. f) Penning et al. *J. Med. Chem.* 1995, *38*, 858-868. g) Cristalli et al. *J. Med. Chem.* 1995, *38*, 1462-1472. h: numbers in parentheses indicate IC_{50} in AT1 subtype assay of series **4**. i: numbers in parentheses indicate IC_{50} in AT2 subtype assay of series **3**.

	Assay	IC_{50} or K_i (nm)		Ref.
1a	Binding to Endothelin A Receptor	400		a
1b	Binding to Endothelin A Receptor	170		a
2a	Inhibition of DNA fragmentation by Topoisomerase I	28		b
2b	Inhibition of DNA fragmentation by Topoisomerase I	143		b
3a	Binding to AT2 subtype of Angiotensin II Receptor	17	(0.45)[h]	c
3b	Binding to AT2 subtype of Angiotensin II Receptor	173	(31)[h]	c
4a	Binding to AT1 subtype of Angiotensin II Receptor	0.85		d

4b	Binding to AT1 subtype of Angiotensin II Receptor	1.4	d
4c	Binding to AT1 subtype of Angiotensin II Receptor	1.2 (23,000)[i]	d
5a	Inhibition of Prolylendopeptidase Protease Activity	5	e
5b	Inhibition of Prolylendopeptidase Protease Activity	10.3	e
6a	Binding to Leukotriene B4 Receptor	320	f
6b	Binding to Leukotriene B4 Receptor	3.2	f
7a	Binding to A2A Type Adenosine Receptor	6.3	g
7b	Binding to A2A Type Adenosine Receptor	41.3	g
8a	none	--	
8b	none	--	
8c	none	--	
8d	none	--	
8e	none	--	

The bulk of these molecules have previously been used as part of an in-depth study validating molecular descriptor approaches for the prediction of molecular diversity *within* compound classes. This is a more stringent discrimination than the base criterion sought for the QCSD model, which seeks at a minimum to show accurate diversity prediction *between* compound classes.

Conformations of all 20 test molecules were "quantized" and then mapped onto the basis set of $1.1 * 10^{14}$ theoretical surfaces. Complementary surfaces are tabulated for each molecule in Table 4.

Table 4: Tabulation of surface shapes and total number of theoretical target surfaces complementary to each molecule in Fig. 5.

	Complementary surface shapes	Complementary surfaces (shape plus functionality)
1a	376	16,127,687
1b	379	9,086,768
2a	27	545,584
2b	27	416,210
3a	414	4,970,816
3b	315	813,024

4a	487	4,542,463
4b	479	12,388,826
4c	482	7,595,982
5a	337	2,080,523
5b	374	1,966,837
6a	220	192,067
6b	186	153,436
7a	362	298,927
7b	269	22,367
8a	45	5,561,654
8b	41	3,959,678
8c	333	17,324,247
8d	64	2,059,546
8e	87	1,343,811
	average:	4,572,523

There are many ways to analyze the resulting set of complementarity mappings. Because in this case individual molecule comparisons were desired, each of the 20 mappings was compared pairwise for a total of 190 data points. Mappings were scored in similarity from 0 to 1000 based on a function of the number of theoretical surfaces in common:

Score = SS * FS = ShapeScore * FunctionalityScore

$$SS = 100 * \frac{\text{\# theoretical target surface shapes common to A \& B}}{\text{total \# surface shapes complementary either to A or to B}}$$

$$FS = 10 * \left(\frac{\text{\# theoretical target surface shapes common to A \& B with at least 1 set of 4 common functionalities)}}{\text{total \# theoretical target surface shapes common to A \& B}} \right)^{\Phi}$$

The first term in this equation gives a percentage measure (0-100) of shape similarity between molecules A and B, while the second term gives a measure from 0-10 of functional similarity per given shape overlap. The complete scores are detailed in Table 5 contained in Figure 22. Using this scoring system, the maximum score obtainable by very rigid, structurally similar molecules is 1000. However, many molecules can only be sampled by an examination of up to 100 low energy conformations (an average molecule w/5+

rotatable bonds will have at least $3^5 = 243$ conformations). Thus, for most molecules with more than 100 accessible conformations, similarity scores between 0-100 are observed. The scoring constant Φ in the equation above adjusts the influence of functionality on scoring. A value of 0.33 was found to be optimal (as discussed below), meaning that shape is the dominant criterion in our measure of diversity. A pseudocode description of an algorithm for determining either the similarity of two molecules or the similarity of two libraries of molecular structures is set forth in Appendix D.

Fig. 6 shows a plot of all 190 pairings ranked by similarity score. Circles show "heterogeneous" pairs of expected dissimilarity (e.g. **2a**, **6b**), while squares show "homogeneous" pairs of expected similarity (e.g. **2a**, **2b**). Clearly, the QCSD model ranks homogeneous pairs almost exclusively higher than heterogeneous pairs; all 15 pharmacologically similar pairs fell within the top 20 scores out of 190. All homogeneous scores were ranked above 25, while the median score in this experiment was 2.8, showing good "signal to noise." The QCSD model is thus a valid predictor of target binding similarity among these molecules.

The pairings also reveal further validation. As might be expected from their relative rigidity (low number of accessible conformations) and structural similarity, the highest scoring pairs are **2a/2b**, **8a/8b**, and **8d/8e**. Furthermore, examination of the pairings of **8c** with **8a,b,d,e** (triangles in Fig. 6, yellow in Table 5) yields scores that are within the top 20% of the pairing experiment but which are generally lower that the "homogeneous" pairs. This makes sense from a target-binding point of view, considering that one face of **8c** contains a large molecular difference (an extra phenyl substituent). To the extent that this face is not involved in complementarity to a target surface, the molecules are similar; to the extent that this face must be complementary for binding to occur, the molecules are quite different. Fig. 7 shows one such case of a surface common to both **8a** and **8c**; the protruding phenyl substituent plays no role in complementarity.

As a rule, there is close similarity of shape and functionality for molecules which score 25 or higher. In addition to the pairings that one would expect, several other pairs scored between 25-35. When these molecules were examined by molecular modeling, significant overlaps were found, suggesting that these high scores are not just "noise" in the QCSD model. Fig. 8 depicts one such case between **1a** and **5a**; conformations of **1a** and **5a** are displayed that were found in the QCSD model to be complementary to the same surface (Fig. 8A, 8B). 3D overlays (Fig. 8C) confirm correlation of general shape and 4 points of functionality, although they also make clear the limits of resolution of complementarity information using 4.24 Å units. As can be seen from Fig. 8, the surface in question can detect general shape and functional similarity, but by does not provide a basis to predict atom-for-atom overlap between molecules.

A final result comes from examination of the QCSD model's rankings of sets **3** and **4**. While set **4** is known to bind exclusively to the AT1 subtype of the Angiotensin II receptor, set **3** is known to bind to both the AT1 subtype and the AT2 subtype. While the QCSD model found high similarity within sets **3** (score = 27) and **4** (avg. score = 54), it found an average similarity of 6.9 between **3a** and set **4** and an average similarity of 3.3 between **3b** and set **4** (diamonds in Fig. 6, green in Table 5 contained in Figure 22). Based on the QCSD model, one would therefore conclude that while sets **3** and **4** share a limited number of complementary theoretical surfaces, they are dissimilar with respect to the majority of theoretical target surfaces. This is in fact the case with the AT2 subtype of the Angiotensin II receptor, to which **4c** binds 50,000 times more poorly than **3a** (see Table 3).

Advantages of the QCSD model

Having validated the basis set used for the QCSD model in the classification of molecular diversity, it must be noted that other models may do as well or better in detecting target binding similarity/ dissimilarity between molecules. For instance, Fig. 9 and Table 6, contained in Figure 23, show the same set of 20 molecules ranked by Tanimoto similarity of standard 2D UNITY fingerprints (see discussion below). The data demonstrate that the 2D model is

equally capable of predicting pharmacologically similar pairs; UNITY ranks similarity between AT1 and AT2 subtype binders much higher than our QCSD model, although it finds unusually high similarity between **8a** and **8c**. In general, such 2D fingerprint descriptors have been found effective in clustering pharmacologically similar compounds, and are widely used in determining molecular diversity of existing structures.

Among the advantages of the QCSD model is the value of its negative information: The QCSD model determines not only diversity of existing structures, but also the structure of non-existing diversity. Given theoretical surface shapes for which no complements exist in a general screening library, QSCD allows the design of molecules to fill the given diversity void.

As stipulated in its formulation, the QSCD basis set is created through a reversible process. Although some information resolution may be lost in fixing the parameters of a cube's size and functional scope, information content is retained in either direction. Just as a single molecular conformation and orientation corresponds to a defined pattern in QSCD space, likewise, a single point in QSCD space (within the limits of volume V, resolution R, and N sites of functionality P_m) corresponds to a unique 3D shape with a defined 3D array of functionality. Given any starting set of molecules, unoccupied points in QSCD space directly define the molecular shapes and functionalities which those molecules do not cover. Thus, a set of detailed 3D molecular templates (at the resolution of the QSCD model used) is immediately available for the creation of novel molecules.

As an example, Fig. 10 shows a plot of all of the theoretical surface shapes covered by all of the conformations of all of the molecules used in the example implementation (see Fig. 5). The total volume of the cube in Fig. 10 encompasses all 49,268,918 theoretical surface shapes as listed in Table 1. As can be seen from the plot and two expanded points, many theoretical surface shapes are "unfilled" by the set of compounds shown in Fig. 5. Thus, in searching for molecules or libraries to enhance the diversity of the given set of compounds, the chemist is presented with a set of actual 3D templates into which

new compound libraries may be designed. In comparison, although mapping the same set of compounds in a "non-reversible" diversity space would also display a set of coordinates to which the molecules map, there would be no way to visualize the 3D shape of any point that was not filled by one of the compounds in the set. Using BCUT values for example (Fig. 11), the coordinates specified for an unfilled point leave the chemist with a set of normalized eigenvalues. While these may give an idea of relative abundance of a given functionality (e.g. H-Bond Donor) at this point in diversity space, the coordinates give no hint of what shape or class of molecules might fill that diversity void.

The above example shows how QSCD is a reversible diversity model with respect to molecular shape. Within a given surface shape in QSCD, there are many combinations of functionality leading to many different theoretical surfaces. If a given library fills only a portion of theoretical surfaces of a given surface shape, by following the same process outlined above and in Fig. 10, unfilled surfaces of specific shape and functionality may be identified and filled with complementary libraries. By using data-mining algorithms to analyze and intersect the shape and functionality of unfilled surfaces, a minimal set of "missing" 3D combinatorial templates can be deduced from the QSCD mapping of a given set of general screening compounds. These templates represent the smallest number of combinatorial syntheses which need to be executed in order to fill out the diversity of the set of screening compounds. One such template is depicted in Fig. 12. In conjunction with the efficiency of core-based combinatorial chemistry, QSCD makes possible the contemplation of a "complete" library of screening molecules at a given resolution. The model thus offers a theoretical and practical answer to the problem of generating lead structures for genomic targets of unknown structure and function.

Technical details of an implementation

For the example discussed above, molecular conformations were generated with Multisearch in Sybyl (version 6.5, Tripos Inc, 1699 S. Hanley Rd., St. Louis, MO, 63144) on an R10000 Silicon Graphics workstation. Conformations were subsequently sorted by energy and conformations within 10

kcal of the lowest energy were accepted. Overlay plots of molecules (Fig. **8B**) were also generated using Sybyl.

UNITY 2D fingerprints (Unity 4.0, Tripos Inc, 1699 S. Hanley Rd., St. Louis, MO, 63144) were generated on an R10000 Silicon Graphics workstation. Pairwise Tanimoto coefficients were computed as described by Dixon and Koehler.

QSCD software for molecule quantization, mapping of Q-files, and surface complementarity display was developed using the Java programming language (JDK 1.2) and the Java3D graphics API (version 1.1) on Intel-based workstations. Theoretical target surfaces were stored and indexed using an Oracle 7.3.3 database.

Parameters for theoretical target surface generation/molecular quantization and parameters for complementarity mapping/scoring were alternately optimized in three successive rounds as described below.

The parameters used for theoretical target surface generation and the closely related parameters for quantization of small molecules into quantized files (Q-files) were optimized in the context of the algorithms mentioned above. Parameters were iteratively optimized by varying a given parameter and then quantizing training molecules other than those in Fig. 5. Training molecules used were taken from in house structures and two published SAR sets. Concomitant with molecular quantization, an enumerated set of theoretical target surfaces was created with corresponding parameters. Using the current optimized complementarity/scoring parameters, molecules were then mapped to theoretical target surfaces and all diversity pairing scores generated as described in the text. Parameters were chosen which accurately predicted known homogeneous/heterogeneous pairs and which maximized "signal to noise" of homogeneous scores over heterogeneous scores.

The parameters used for mapping/scoring molecular conformations to theoretical target surfaces were optimized in the context of the algorithm stated above. Parameters were iteratively optimized by varying a given parameter and then mapping a constant set of training molecules (see above) to a constant set of

theoretical target surfaces, using the most current surface generation and quantization parameters. Diversity pairing scores were generated for all training molecules, and parameters were chosen which accurately predicted known homogeneous/heterogeneous pairs and which maximized "signal to noise" of homogeneous scores over heterogeneous scores.

As mentioned earlier, for a given conformation-to-surface shape fit to be accepted, the minimum overlap requirement was set to either 9 quanta or N-2 quanta of a conformation of N quanta. This range allows large conformations to fit partially into a theoretical surface (protruding volume must be at the mouth of the surface) while also allowing smaller conformations to be considered for complementarity. It excludes large conformations which do not overlap at least 9 quanta.

Approximate computational speeds of typical QSCD operations are as follows on a single Pentium III 500 MHz workstation: Generation of the basis set of theoretical target surface used in the study required 17 min.; this data was stored for access by subsequent QSCD functions. Quantization of 100 conformations of a given molecule into 100 Q-files required 250 seconds. Complementarity mapping of 100 Q-files onto the basis set of theoretical target surfaces used in the study required 40 seconds.

Algorithm for Designing Molecules for Unfilled Target Surfaces

The following algorithm could be used to design novel molecules based on complementarity to unfilled theoretical target surfaces that are not complementary to any existing molecular conformations.

1) Existing molecules are quantized and those negative space cube target surfaces to which their conformations are complementary are identified. (see Appnedixes A, B, and C)

2) For a given set of existing molecules and a desired set of theoretical target surfaces with given shapes and functionalities, those theoretical target surfaces are identified to which no existing molecular conformations are complementary. Novel molecules are designed to be complementary to the above identified theoretical target surfaces as follows:

   a) Let the set of those theoretical target surfaces to which no existing molecular conformations are complementary = T.

b) `Cluster T into subsets T1-Tn such that for each Ti there is a central set of negative space cubes of given functionality Ci such that all target surfaces in Ti can be created by adding up to E additional negative space cubes of given functionality to Ci at up to each of F points of attachment on Ci where each point of attachment Aj (j = 1–F) is a single face of Ci to which zero, one or a set of up to E negative space cubes may be added. (See Fig. 24)

c) For each Ci, design a core molecule Mi that fills but does not extend beyond the space defined by Ci within a tolerance limit TOL, the core molecule furthermore complementary to the functionality of Ci, and the core molecule furthermore containing at least one combinatorial site (defined as a reactive site that can be further functionalized and/or extended in a combinatorial step under given chemical conditions) that can project potential combinatorial building blocks through at least one plane Aj. (See Fig. 25)

d) Computationally enumerate the combinatorial library L(Mi,B) defined by Mi and a set of building blocks B that are each no larger in volume than E negative space cubes (See Fig. 26).

e) Quantize n_c conformations of each molecule in L(Mi,B) and determine the set W of all theoretical surfaces to which any conformation of any molecule in L(Mi,B) is complementary (see Appendixes B and C).

f) Compute the set of target surfaces $M = W \cap Tn$

g) If M contains an acceptable number of novel surfaces as defined by the user, then chemically synthesize the actual library L(Mi,B). Otherwise, choose a new Mi in step 3 for the given Ci and repeat until conditions in this step are met.

h) Move on to the next Ci in step c.

Protein Diversity

An extension of any diversity model based on an absolute frame of reference is that the same basis set may be used to classify actual proteins. By mapping onto the QSCD basis set all surfaces of volume V of a known protein, actual proteins can be compared and classified by their 3D binding sites. In addition to providing a diversity map of known protein binding sites within the universe of all theoretical protein surfaces under given parameters, the theoretical

surfaces of QSCD may thus be used to correlate protein classes to complementary molecular core structures.

Appendix E describes an algorithm for quantization of protein surfaces. Appendix F describes an algorithm for comparing protein surfaces to determine a degree of similarity or dissimilarity. The following algorithm generates a set of files T of quantized protein binding surfaces which together represent the available surface of a given protein binding site.

Algorithm T(protein binding site, H, R, V, TolA, TolB, Transinc, Rot, Rotvar, P_M)

Protein binding site: see 1. below
H: resolution of scanning grid in angstroms, H<=R
R: resolution of cube = length of cube side in angstroms
V: total volume of each surface in cubic angstroms
TolA: tolerance (%) of atomic radii
TolB: tolerance (%) of cube volume which must intersect convex hull
Transinc: translational increment in angstroms
Rot: # rotations (odd integer)
Rotvar: rotational variance (%)
P_M: For each of V/R^3 cubes used, any of M types of chemical functionality

1. Take a protein binding site file (obtained by any number of commercially available methods, such as running a "Connolly surface search" on a standard "PDB file" (Michael L. Connolly, 1259 El Camino Real, #184, Menlo Park, CA 94025) which minimally consists of:

   a) A calculated probable electron density surface of the binding site
   b) A list of all known atom types in the molecule with their coordinates and atomic radii
   c) A list of known connectivities of all atoms with the type of bond connecting each atom

2. Overlay flexible 2D square grid(s) of grid size HxH angstroms over all calculated probable electron density surfaces in the protein binding site

3. Calculate the convex hull of the set of points defined by the points of each grid nexus

4. Examine each grid nexus one at a time

5. At a given grid nexus place a cube with the center of one face tangent to the probable electron density surface of the protein binding site

6. If the cube contains a protein atom coordinate or any atomic radii of protein atoms protrude into the trial cube by more than Tol % of their atomic radius, then step 7., otherwise step 8.

7. Translate the cube Transinc angstroms away from the grid nexus on the probable electron density surface while keeping the center of the cube on a line perpendicular to the probable electron density surface at the the grid nexus. If the cube is now R angstroms away from the grid nexus, take the next nexus in 4., otherwise return to step 6.

8. Designate the cube as being a set cube with 5 unchecked faces and 1 checked face (the checked face being that which faces the grid nexus being examined). Designate the initial frame of reference to be the X, Y, and Z axes co-linear with the sides of the cube.

9. Rotate the unchecked cube about its center in all combinations (total of Rot^3 combinations) of the following units:

   a) X rotations: any one of Rot units (degrees) from - Rotvar * 90/2 to +Rotvar * 90/2 by Rotvar * 90/Rot
   b) Y rotations: any one of Rot units (degrees) from - Rotvar * 90/2 to +Rotvar * 90/2 by Rotvar * 90/Rot
   c) Z rotations: any one of Rot units (degrees) from - Rotvar * 90/2 to +Rotvar * 90/2 by Rotvar * 90/Rot

10. For a given rotated system from 9., place face contiguous trial cubes of side length R at all unchecked cube face(s), not to exceed the nearest integer to V/R^3 total cubes. If addition of a new trial cube would exceed the nearest integer to V/R^3 total cubes, proceed directly to 11 without adding further trial cubes

11. All unchecked cube faces (not including cube faces on trial cubes) become checked cube faces.

12. If a trial cube contains a protein atom coordinate, or any atomic radii of protein atoms protrude into the trial cube by more than TolA of their atomic radius, or if the cube does not intersect with a volume of the convex hull equal to at least R^3 * TolB (see 3. above), then the trial cube is removed. Otherwise the trial cube becomes a set cube (with unchecked faces).

13. If the total number of set cubes becomes = the nearest integer to V/R^3 (yielding a resulting cube combination of the nearest integer to V/R^3 cubes), then go to step 15.

14. If no trial cubes in step 12 became set cubes and there are no unchecked cube faces remaining then start with the next rotated cube in step 9. If no rotations produce resulting cube combinations in step 13. then start with the next translation in step 7.

15. Ignore any remaining rotations from step 9. Designate all cubes as negative space cubes fully enclosed except as detailed below. Designate the layer of cubes which is

a) perpendicular to the line perpendicular to the probable electron density surface at the grid nexus being examined
b) farthest from the grid nexus

as negative space cubes which are open at their faces farthest from the grid nexus and perpendicular to the line perpendicular to the probable electron density surface at the grid nexus.

16. Based on the types of proximal atoms in the protein surrounding each negative space cube, and based upon the bonds which these proximal atoms form and the other atoms to which these proximal atoms are bonded, ascribe to each negative space cube one and only one of M types of chemical functionality.

Types of functionality M may include but are not limited to:

Acidic regions,
Basic regions,
regions of formal charge +1,
regions of formal charge –1,
regions of partial charge between +0.5 and +1,
regions of partial charge between -0.5 and -1,
regions of partial charge between 0 and +0.5,
regions of partial charge between 0 and -0.5,
hydrophobic regions,
polarizable regions,
hydrogen bond donating regions
hydrogen bond accepting regions
hydrogen bond donating/accepting regions

17. Yield a "quantized" protein binding surface in terms of the nearest integer to V/R^3 negative space cubes of side length R with any one of M types of functionality per negative space cube.

18. Return to step 4. for each grid nexus

19. Compare all quantized surfaces from 17. and remove any which are identical

20. Yield a set of files T of quantized protein binding surfaces which together represent the available surfaces of the given protein binding site

Comparing Molecules to Proteins

Appendix G describes an algorithm for determining the complementarity of a library of molecules to a set of protein surfaces.

Algorithm for Designing Molecules for a Set of Protein Target Surfaces

In another procedure, novel molecules (for testing as ligands for proteins) can be designed based on complementarity to negative space cube targets to which a set of protein pockets map. The following outline describes the steps for doing so:

1) A set of existing protein pockets is quantized and those negative space cube target surfaces to which their quantizations map are identified. (See Appendixes B and C.)

2) Novel molecules are designed to be complementary to the above identified set of target surfaces as follows:

   a) Let the set of target surfaces = T.

   b) Cluster T into subsets T1-Tn such that for each Ti there is a central set of negative space cubes of given functionality Ci such that all target surfaces in Ti can be created by adding up to E additional negative space cubes of given functionality to Ci at up to each of F points of attachment on Ci where each point of attachment Aj (j = 1–F) is a single face of Ci to which zero, one or a set of up to E negative space cubes may be added. (See Fig. 27)

   c) For each Ci, design a core molecule Mi that fills but does not extend beyond the space defined by Ci within a tolerance limit TOL, the core molecule furthermore complementary to the functionality of Ci, and the core molecule furthermore containing at least one combinatorial site (defined as a reactive site that can be further functionalized and/or extended in a combinatorial step under given chemical conditions) that can project potential combinatorial building blocks through at least one plane Aj. (See Fig. 28)

d) Computationally enumerate the combinatorial library L(Mi,B) defined by Mi and a set of building blocks B that are each no larger in volume than E negative space cubes (See Fig. 29).

e) Quantize n_c conformations of each molecule in L(Mi,B) and determine the set W of all theoretical surfaces to which any conformation of any molecule in L(Mi,B) is complementary (See Appendixes B and C.)

f) Compute the set of target surfaces $M = W \cap Tn$

g) If M contains an acceptable number of novel surfaces as defined by the user, then chemically synthesize the actual library L(Mi,B). Otherwise, choose a new Mi in step 3 for the given Ci and repeat until conditions in step 7 are met.

h) Move on to the next Ci in step 3.

Parameter Values

Appendix H contains example parameter values useful in connection with the algorithms described in Appendixes A through G.

Further Extensions

As mentioned, a 4.24 Å cube was found to be the largest predictive unit size of diversity measure for our criteria of designing general screening libraries. For example, both 4.48 and 4.00 Å units gave poorer prediction of homogeneous/heterogeneous pairs than the pairings of Fig. 6 (4.24 Å units). This is likely due to the fact that most organic small molecules are themselves quantized by a limited basis set: the VDW radii of H, C, N, O and a few other atoms (see for example Fig.1). If there is no constraint on size of cubic units, however (i.e., if there is no attempt to maximize orthogonality of theoretical target surfaces), other unit measures of diversity can be found. A unit of 2.12 Å should also provide effective diversity information but at a much higher resolution. Such a "high-resolution" adaptation of QSCD brings with it numerical (and thus computational) challenges. 112 negative space cubes (14 x 8) are now required at the upper limit of theoretical target surface size, translating to exponentially greater numbers of theoretical target surfaces and, depending on

the stringency of fitting parameters, correspondingly greater numbers of surface fits per molecule as in Table 4. At this resolution, the assumption of no occlusions in theoretical target surfaces becomes less valid, and removal of this assumption increases computational complexity further.

A corollary of any absolute diversity model is a prediction of the total "size" of diversity space in terms of unique molecular points. In other words, what is the minimum set of molecules needed to fully cover a given diversity space. This calculation is dependent on two factors: the resolution stipulated in the model (e.g., what amount of molecular change is recognized as different) and the maximum values of each dimension of the model's basis axes. In the model of QSCD discussed above, resolution is fixed by cubic units of 4.24 Å, and maximum values are fixed at 14 units (molecular volume of 1070 cubic Å) and 4 points of 7 types of molecular property characteristics. As describe above, the result is a set of $1.1 * 10^{14}$ unique molecular points. Since, using the parameters of this study, an average molecule covers 4.6 million of the unique molecular points bounded by QSCD space (Table 4), the model predicts a minimum $1.1 * 10^{14} / 4.6 * 10^{6}$ = 24 million molecules would be necessary to completely cover diversity space.

We estimate that an average complementary molecule in the context of the QSCD model has a ΔG of complementarity on the order of -11 kcal (Table 7).

Table 7. Summation of binding energies for an interaction of an average complementary molecule/theoretical target surface pair in the context of the QSCD model used herein. An average molecule is assumed to have a buried volume of 12 cubic quanta (= 915 cubic Å at 4.24 Å resolution), 36 exposed faces (4.24 Å square), 21 non-polar exposed faces (60%), 10 rotatable bonds, 4 points of complementary electrostatic/VDW potential, and a conformational energy within 2 kcal/mol of ground state. An average complementary theoretical target surface is also assumed to have 60% non-polar exposed faces. Constants used in the table are taken from Ajay and Murcko.

Energetic contribution	Average ΔG

translational/vibrational entropic loss (constant)	+9 kcal/mol
constant +0.7 kcal/mol (RTln3) per rotatable bond; assume rigid theoretical target surface	+7 kcal/mol
$\Delta\Delta G$ conformation from ground state	+2 kcal/mol
constant -0.03 kcal/mol per square Å non-polar buried surface = -0.54 kcal/mol per 4.24 Å square non-polar buried face; total 21 molecular faces + 21 theoretical surface faces	-23 kcal/mol
total interaction from Table 2 (four complementary points)	-6.0 kcal/mol
sum of binding energies	-11 kcal/mol
binding affinity to nearest integer (÷1.363)	**10^{-8} = 10 nanomolar**

In other words, the resolution used to calculate diversity translates roughly to nanomolar binding conditions for an average molecule/target surface pair. Given that some 24 million molecules are needed to completely cover diversity space under these conditions, a general screening library guaranteed to contain at least one nanomolar binder to any given target of interest would thus number at least 24 million molecules. This is a large number and will be attenuated by the fact that some molecules have significantly more than 100 conformations available to them. However, the QSCD model suggests that if, in the near future, combinatorial chemistry and high-throughput screening are to generate initial hits primarily in the nanomolar rather than micromolar range, then the field must continue to focus its efforts on the development of numerically competent synthesis and screening technologies.

By defining molecules in terms of their complementarity to a fully enumerated set of theoretical protein surfaces under given parameters, and by defining actual protein surfaces in terms of their similarity to the same set of theoretical protein surfaces, the model allows:

1. Numerical prediction of protein surface diversity (how many different types of possible protein surfaces exist under a given set of parameters).
2. Numerical prediction of how many and what type of molecules would be necessary to create a "universal molecular library" (a library which contains at least one complement [of a given score] to any given protein surface).
3. Comparison of the similarity or difference of molecules or sets of molecules based on their complementarity to theoretical protein surfaces. The frame of reference for such comparisons is fixed no matter how many or what types of molecules are involved.
4. Numerical prediction of the percent of protein surface diversity to which a given set of molecules is complementary.
5. Comparison of the similarity or difference of actual protein surfaces or sets of surfaces based on their similarity to theoretical protein surfaces. The frame of reference for such comparisons is fixed no matter how many or what types of protein surfaces are involved.
6. Numerical prediction of the percent of protein surface diversity to which a given set of actual protein surfaces is similar.
7. Prediction of the actual protein surfaces to which a given molecule is complementary.
8. Prediction of how many and what type of molecules would be necessary to create a "universal molecular library" against a given set of actual protein surfaces (a library which contains at least one complement [of a given score] to each actual protein surface).

This application discloses information from an article which has been accepted for publication in the peer-reviewed Journal of Medicinal Chemistry and is currently scheduled for publication in the May 18th issue. The article, listed by the Journal of Medicinal Chemistry as JM990504B, is titled "Quantized

surface complementarity diversity (QSCD): A model based on small molecule-target complementarity," and is incorporated herein by reference.

Other implementations are within the scope of the following claims.

Appendix A

Theoretical Surfaces

A *surface opening* O is a set of lattice squares in $\mathbb{Z}^2$ denoted by their corners:

$$O = \{(x_i, y_i) \in \mathbb{Z}^2, i = 1 \ldots m\}$$

By definition, a surface opening must be connected. For connectivity purposes, two points $p, q \in \mathbb{Z}^2$ are neighbors iff

$$|p_x - q_x| + |p_y - q_y| = 1$$

The area of a surface opening is defined to be the number of lattice squares it contains. Surface openings are considered to be unique upto translations and rotations of the x-y plane.

A *surface shape* S is a set of "negative space" cubes represented as lattice cubes in $\mathbb{Z}^3$ denoted by their corners:

$$S = \{(x_i, y_i, z_i) \in \mathbb{Z}^3, i = 1 \ldots n\}$$

By definition, a surface shape must satisfy the following conditions:

- All cubes are below the x-y plane. That is, if $(x, y, z) \in S$, then $z \leq 0$.
- The surface shape is connected. For connectivity purposes, two cubes $p, q \in \mathbb{Z}^3$ are neighbors iff

$$|p_x - q_x| + |p_y - q_y| + |p_z - q_z| = 1$$

- There are no occlusions along the z-axis. That is, if $(x, y, z) \in S$ and $z < 0$, then $(x, y, z + 1) \in S$.

The volume of a surface shape is defined to be the number of lattice cubes it contains. Surface shapes are considered to be unique up to translations and rotations of the x-y plane.

Every surface shape specifies a unique opening:

$$\text{opening}(S) = \{(x, y) | (x, y, z) \in S\}$$

APPENDIX A. THEORETICAL SURFACES

Further, given a surface opening O and a function $d : O \rightarrow \mathbb{N}$, a unique surface shape with the given opening is defined:

$$\text{shape}(O, d) = \{(x, y, z) | (x, y) \in O, d(x, y) > -z\}$$

The function d specifies the "depth" of the surface shape at each opening point. Sample surface shapes and openings can be seen in Figure 13.

A *theoretical surface* consists of a surface shape where the cubes in the surface shape are each associated with functionality. The set $\mathcal{F}$ of seven specific types of characteristic functionality is used:

- $\mathcal{F}_1$: Potential Negative Charge
- $\mathcal{F}_2$: Potential Positive Charge
- $\mathcal{F}_3$: Hydrogen Bond Donor/Acceptor
- $\mathcal{F}_4$: Hydrogen Bond Donor
- $\mathcal{F}_5$: Hydrogen Bond Acceptor
- $\mathcal{F}_6$: Polarizable
- $\mathcal{F}_7$: Hydrophobic
- $\mathcal{F}_8$: Slightly Hydrophobic

A functionality map $f : S \rightarrow \mathcal{F}$ defines the assignment. By default, all cubes are assigned functionality $\mathcal{F}_8$, and all possibilties are considered where upto n_f of the cubes are given one of the functionalities $\mathcal{F}_1$-$\mathcal{F}_7$.

Generating all surface shapes is accomplished by the following steps:

1. Generate all surface openings (SURFACEOPENINGS).
2. Filter the surface openings to remove openings that are unlikely to resemble surfaces found in nature (OPENINGFILTER).
3. From the set of filtered openings, generate all surfaces shapes whose associated openings are in the set (SURFACESHAPES).
4. Filter the set of surfaces shapes to remove shapes unlikely to resemble surfaces found in nature (SHAPEFILTER).
5. From the set of filtered surface shapes, add all possible combinations of functionality to define a set of theoretical surfaces (FUNCTIONALIZESURFACE).

APPENDIX A. THEORETICAL SURFACES

Algorithm A.1 SURFACEOPENINGS(A): calculate all surface openings with area less than or equal to A.

```
1: define N_1 to be a set containing the only the surface opening with a single square
   at (0, 0)
2: for i ← 2 to A do
3:    N_i ← ∅ {N_i will ultimately contain all openings of area i}
4:    for all O ∈ N_{i-1} do
5:       define P to the set of all possible openings obtained by adding a single square
         to O adjacent to a square already present in O
6:       for all P ∈ P do
7:          if no rotation or translation of P is present in N_i then
8:             add P to N_i
9:          end if
10:      end for
11:   end for
12: end for
13: O ← ⋃_{i=1}^{A} N_i
14: return (O)
```

APPENDIX A. THEORETICAL SURFACES

Algorithm A.2 OPENINGFILTER($\mathcal{O}, A_t, M_{nc}, M_c$): filter a set of surface openings $\mathcal{O}$ using area-threshold parameter A_t, max-non-central parameter M_{nc}, and max-contiguous parameter M_c.

```
1: Ô ← ∅
2: for all O ∈ 𝒪 do
3:    Ô ← O
      {delete from Ô all lattice squares with exactly one neighbor that has 2 or more
      other neighbors}
4:    for all (x, y) ∈ Ô do
5:       N ← Ô ∩ {(x − 1, y), (x + 1, y), (x, y − 1), (x, y + 1)}
6:       if |N| = 1 then
7:          define (x̂, ŷ) to be the single element in N
8:          N̂ ← Ô ∩ {(x̂ − 1, ŷ), (x̂ + 1, ŷ), (x̂, ŷ − 1), (x̂, ŷ + 1)}
9:          if |N̂| > 2 then
10:            remove (x, y) from Ô
11:         end if
12:      end if
13:   end for
      {delete from Ô all 2x2 blocks of lattice squares in O}
14:   for all (x, y) ∈ Ô do
15:      if (x + 1, y) ∈ O and (x, y + 1) ∈ O and (x + 1, y + 1) ∈ O then
16:         remove (x, y), (x + 1, y), (x, y + 1), (x + 1, y + 1) from Ô
17:      end if
18:   end for
19:   define c to be the number of squares in the largest connected component left in
      Ô
20:   if area(Ô) ≤ M_nc and (area(O) ≤ A_t or c ≤ M_c) then
21:      add Ô to 𝒪̂
22:   end if
23: end for
24: return (𝒪̂)
```

APPENDIX A. THEORETICAL SURFACES

Algorithm A.3 SURFACESHAPES($\mathcal{O}, V$): calculate all surface shapes with openings in the set $\mathcal{O}$ and volume less than or equal to V.

1: **for all** $O \in \mathcal{O}$ **do**
2: define $d : O \rightarrow \mathbb{N}$ such that $d(x, y) = 1$
3: $\mathcal{S}_O \leftarrow \emptyset$ {$\mathcal{S}_O$ will contain all surfaces with opening O}
4: $v \leftarrow \text{area}(O)$
5: **while** $v \leq V$ **do**
6: $S \leftarrow \text{shape}(O, d)$
7: **if** $\mathcal{S}_O$ contains no translations or rotations of S **then**
8: add S to $\mathcal{S}_O$
9: **end if**
10: **for** $x \leftarrow -\text{area}(O)$ **to** $\text{area}(O), y \leftarrow -\text{area}(O)$ **to** $\text{area}(O)$ **do**
11: **if** $(x, y) \in O$ **and** $v + 1 \leq V$ **then**
12: $v \leftarrow v + 1, d(x, y) \leftarrow d(x, y) + 1$
13: **goto step** 5
14: **else**
15: $v \leftarrow v - d(x, y) + 1, d(x, y) \leftarrow 1$
16: **end if**
17: **end for**
18: **goto step** 1 {no more surfaces left with this opening}
19: **end while**
20: **end for**
21: $\mathcal{U} \leftarrow \bigcup_{O \in \mathcal{O}} \mathcal{S}_O$
22: **return** ($\mathcal{U}$)

Algorithm A.4 SHAPEFILTER($\mathcal{S}, M_e$): filter a set of surfaces $\mathcal{S}$ using max-extrusion parameter M_e.

1: $\hat{\mathcal{S}} \leftarrow \emptyset$
2: **for all** $S \in \mathcal{S}$ **do**
3: $O \leftarrow \text{opening}(S)$
4: $d \leftarrow \text{depth}(S)$ {corresponding depth function}
5: **for all** $(x, y) \in O$ **do**
6: **for all** $(\hat{x}, \hat{y}) \in \{(x - 1, y), (x + 1, y), (x, y - 1), (x, y + 1)\}$ **do**
7: **if** $(\hat{x}, \hat{y}) \in O$ **and** $d(\hat{x}, \hat{y}) \geq d(x, y) - M_e$ **then**
8: **goto step** 5
9: **end if**
10: **end for**
11: **goto step** 2 {surface does not qualify}
12: **end for**
13: add S to $\hat{\mathcal{S}}$ {surface does qualify}
14: **end for**
15: **return** ($\hat{\mathcal{S}}$)

APPENDIX A. THEORETICAL SURFACES

Algorithm A.5 FUNCTIONALIZESURFACE(S, n_f): Return the set of theoretical surfaces defined by surface shape S with n_f points of specific functionality attached.

1: $\mathcal{S} \leftarrow \emptyset$ {set of functionalized surfaces }
2: define $\mathcal{V}$ to be the set of all the $\binom{\text{volume}(S)}{n_f}$ selections of n_f cubes from S
3: **for all** $V \in \mathcal{V}$ **do**
4: define $f : S \rightarrow \mathcal{F}$ such that $f(s) = \mathcal{F}_8$ if $s \notin V$, $f(s) = \mathcal{F}_1$ for $s \in V$
5: number the elements of V as $v_1, \ldots, v_{n_f}$
6: **loop**
7: add (S, f) to $\mathcal{S}$
8: **for** $i \leftarrow 1$ **to** n_f **do**
9: **if** $f(v_i) \neq \mathcal{F}_7$ **then**
10: assign $f(v_i)$ to be the next higher functionality
11: **goto step** 6
12: **else**
13: $f(v_i) \leftarrow \mathcal{F}_1$
14: **end if**
15: **end for**
16: **goto step** 3{no more functionality assigments left with the set V of cubes}
17: **end loop**
18: **end for**
19: **return** (S)

Appendix B

Quantization

In the quantization process a molecule is reduced into a representation such that complementarity can be calculated against a set of theoretical target surfaces. Quantization takes place in the following steps:

1. Each atom in the molecule is assigned a functionality based on its type and connectivity.
2. Three dimensional conformations of the molecular structure are generated.
3. Each conformation is converted into "positive space" cubes based on the positions of its atoms.
4. Each cube is assigned a functionality based on the atoms that it contains.

Finally, the quantized form of the molecule is defined to be the set of all selected conformation quantizations with functionalities assigned. The quantization process is summarized in Figure 14.

APPENDIX B. QUANTIZATION

B.1 Atomic Functionality Assignment

Given a molecular structure as a set of atoms $\mathcal{M}$, a map $f : \mathcal{M} \rightarrow \mathcal{F}$ is defined assigning functionalities to all of the atoms. The map is defined by the using a set of rules to match molecular substructures based on extended atom types (as generated by Tripos, for example) and bonding paterns that encapsulate each functionality type. The algorithm keeps track of atoms that are excluded from matching a lower priority rule because they have already been matched in a higher priority rule, where $\mathcal{F}_1$ has the highest priority and $\mathcal{F}_7$ the lowest. Atoms not matching any rule are assigned functionality $\mathcal{F}_7$. No atoms are assigned functionality $\mathcal{F}_8$. The functionality rules used can be seen as follows:

- $\mathcal{F}_1$: Figure 16
- $\mathcal{F}_2$: Figure 17
- $\mathcal{F}_3$: Figure 18
- $\mathcal{F}_4$: Figure 19
- $\mathcal{F}_5$: Figure 20
- $\mathcal{F}_6$: Figure 21

Within each functionality type, the functionality rules are search sequentially in the order listed in Figures 16–21. Figure 15 provides a legend for the symbols used in the functionality rule diagrams.

Algorithm B.1 ATOMFUNCTIONALMAP($\mathcal{M}$): assign functionality to the atoms in molecular structure $\mathcal{M}$ and determine which atoms are excluded from quantization.

```
1: E_c ← ∅ {E_c is the set of atoms that are excluded from consideration. }
2: E_q ← ∅ {E_q is the set of atoms that are excluded from the rest of the quantization
   process.}
3: define f : M → F such that f(a) = F_7 for all atoms a ∈ M.
4: for i ← 1 to 6 do
5:    for each functionality rule for F_i considered in the proper order do
6:       find all matches in M with consideration to E_c
7:       according to the rule, add appropriate atoms to E_c and E_q; and set f to F_i for
         selected atoms
8:    end for
9: end for
10: return (f, E_q)
```

APPENDIX B. QUANTIZATION

B.2 Conformation Generation

A conformation is a mapping $c : \mathcal{M} \rightarrow \mathbb{R}^3$ of the molecular structure into three dimensional space. Using OpenEye Omega software (Open Eye Scientific Software Inc., 335c Winische Way, Santa Fe, NM, 87501), upto n_c representation conformations for each molecule are generated within given rule-based energy parameters.

APPENDIX B. QUANTIZATION

B.3 Base Coordinate Frame

A coordinate frame $T = (R, t) : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is a rigid motion of the space defined by a rotation $R \in \mathrm{SO}_3(\mathbb{R})$ and a translation $t \in \mathbb{R}^3$ that transforms a point p by the rule:

$$T(p) = Rp + t$$

Given a resolution r (the length of a side of a lattice cube) and a coordinate frame, a lattice on $\mathbb{R}^3$ is implicitly defined by

$$q \in \mathbb{Z}^3 \leftrightarrow [rq_x, rq_x + r) \times [rq_y, rq_y + r) \times [rq_z, rq_z + r) \subset \mathbb{R}^3$$

The base coordinate frame is generated from a conformation of molecular structure $\mathcal{M}$. First, a subset $\tilde{\mathcal{M}}$ of the atoms in $\mathcal{M}$ are selected via FRAMEATOMS. These are atoms in or near ring structures close to the center of the conformation. A ring atom is defined to be an atom that contains at least one bond which, if removed, would not result in the molecule being disconnected. If there are an insufficient number of ring atoms, all atoms sufficently close to the center of the conformation are used.

Then, the base coordinate frame is calculated in BASEFRAME. The x-axis of the base coordinate frame is defined to to be the solution to the optimization problem:

$$\max_{\|x\|=1} \sum_{a \in \tilde{\mathcal{M}}} |x^T(c(a) - p)|$$

where

$$p = \frac{1}{|\tilde{\mathcal{M}}|} \sum_{a \in \tilde{\mathcal{M}}} c(a)$$

is the center of the conformation. Due to the non-linear nature of this problem, it is solved using an approximate gradient descent method. Given x, the y-axis of the base coordinate frame is defined to be the solution of a second optimization problem:

$$\max_{\|y\|=1, x^T y=0} \sum_{a \in \tilde{\mathcal{M}}} |y^T(c(a) - p)|$$

Given x and y, z is uniquely specified. The base coordinate frame then simply involves centering the conformation by translating p to the origin, and using the new x, y, and z axes.

APPENDIX B. QUANTIZATION

Algorithm B.2 FRAMEATOMS($\mathcal{M}, c, r_f, r_m, q_r$): select a set of atoms to use for generating the coordinate frame from a molecular structure $\mathcal{M}$ and a conformation c. The following parameters are used: ring factor r_f, ring minimum r_m, radius factor q_r.

$\mathcal{R} \leftarrow \emptyset$ {$\mathcal{R}$ will hold the set of ring atoms }
for all $a \in \mathcal{M}$ **do**
  if a is not a Hydrogen atom **and** a is in a ring **then**
    add a to $\mathcal{R}$
  end if
end for
$p \leftarrow \frac{1}{|\mathcal{M}|} \sum_{a \in \mathcal{M}} c(a)$ {the mathematical center of the conformation }
$r \leftarrow \max_{a \in \mathcal{M}} \|c(a) - p\|$ {the maximum distance from the center of the conformation}
$\mathcal{C} \leftarrow \emptyset$ {a placeholder for atoms in rings we have already examined }
$\mathcal{R}_g \leftarrow \emptyset$ {a set of rings which are close to the center of the conformation }
for all $a \in \mathcal{R}$ **do**
  if $\|c(a) - p\| \leq r q_r$ **and** $a \notin \mathcal{C}$ **then** {if a ring atom is sufficiently close to the center of the conformation, add all members of its ring group }
    $\mathcal{G} \leftarrow$ RINGGROUP($a, \mathcal{R}, r_f$)
    add $\mathcal{G}$ to $\mathcal{R}_g$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{G}$
  end if
end for
$\mathcal{D} \leftarrow \{a \in \mathcal{M}$ such that$\|c(a) - p\| \leq r q_r\}$ {a default set of atoms close to the center, to be used if we haven't collected enough ring atoms }
if $\mathcal{R}_g = \emptyset$ **then** {if there are no ring groups, use the default }
  return ($\mathcal{D}$)
end if
if $\max_{\mathcal{G} \in \mathcal{R}_g} |\mathcal{G}| < r_m$ **then** {if the biggest ring group is smaller that r_m, use the default}
  return ($\mathcal{D}$)
end if
$\mathcal{B}_g \leftarrow \bigcup_{\mathcal{G} \in \mathcal{R}_g, |\mathcal{G}| > 6} \mathcal{G}$ {a set of "big" ring groups, having at least 6 atoms }
$\tilde{\mathcal{M}} \leftarrow \emptyset$
for all $a \in \bigcup_{\mathcal{G} \in \mathcal{B}_g} \mathcal{G}$ **do** {use all "big" ring group atoms and any neighboring atoms}
  add a and all atoms bonded to a to $\tilde{\mathcal{M}}$, if not already present
end for
return ($\tilde{\mathcal{M}}$)

APPENDIX B. QUANTIZATION

Algorithm B.3 RINGGROUP($a, \mathcal{R}, r_f$): calculate the set of atoms that can be reached starting at atom a and crossing over at most r_f atoms that are not in the set of ring atoms $\mathcal{R}$.

```
1: d ← 0
2: T_h ← {a}, T_n ← ∅
3: E ← {a}, G ← {a}
4: while d ≤ r_f do
5:    if T_h = ∅ then
6:       d ← d + 1
7:       T_h ← T_n, T_n ← ∅
8:    else
9:       â ← pop(T_h)
10:      if â is not a Hydrogen atom and â ∉ E then
11:         for all atoms o bonded to â do
12:            if o is not a Hydrogen atom and o ∉ E then
13:               add o to E
14:               if o ∈ R then
15:                  add o to G
16:                  add o to T_h
17:               else
18:                  add o to T_n
19:               end if
20:            end if
21:         end for
22:      end if
23:   end if
24: end while
25: return (G)
```

Algorithm B.4 BASEFRAME($\tilde{\mathcal{M}}, c$): compute the base coordinate frame given a set of atoms $\tilde{\mathcal{M}}$ and a conformation c.

1: $p \leftarrow \frac{1}{|\tilde{\mathcal{M}}|} \sum_{a \in \tilde{\mathcal{M}}} c(a)$
2: define $u : \tilde{\mathcal{M}} \rightarrow \mathbb{R}^3$ by $u(a) = c(a) - p$
3: $d_1 \leftarrow (0, 0, 0)$
4: **for all** $a \in \tilde{\mathcal{M}}$ **do**
5: $v \leftarrow u(a)/\|u(a)\|$
6: **loop**
7: $s \leftarrow \sum_{o \in \tilde{\mathcal{M}}} \text{sign}(v^T u(o)) u(o)$
8: $s \leftarrow s/\|s\|$
9: **for all** $o \in \tilde{\mathcal{M}}$ **do**
10: **if** $\text{sign}(s^T u(o)) \neq 0$ **and** $\text{sign}(s^T u(o)) \neq \text{sign}(v^T u(o))$ **then**
11: $v \leftarrow s$
12: **goto step 6**
13: **end if**
14: **end for**
15: $v \leftarrow s$
16: **goto step 18**
17: **end loop**
18: **if** $\sum_{a \in \tilde{\mathcal{M}}} |v^T u(o)| > \sum_{a \in \tilde{\mathcal{M}}} |d_1^T u(o)|$ **then**
19: $d_1 \leftarrow v$
20: **end if**
21: **end for**
22: $d_2 \leftarrow (0, 0, 0)$
23: **for all** $a \in \tilde{\mathcal{M}}$ **do**
24: $v \leftarrow d_1 \times (u(a) - d_1^T u(a) d_1)$
25: $v \leftarrow v/\|v\|$
26: $s \leftarrow \sum_{o \in \tilde{\mathcal{M}}} \text{sign}(v^T u(o)) u(o)$
27: $s \leftarrow (s - d_1^T s d_1)/\|s - d_1^T s d_1\|$
28: **if** $\text{sign}(s^T u(o)) = \text{sign}(v^T u(o))$ for all $o \in \tilde{\mathcal{M}}$ such that $\text{sign}(s^T u(o)) \neq 0$ **then**
29: **if** $\sum_{a \in \tilde{\mathcal{M}}} |s^T u(o)| > \sum_{a \in \tilde{\mathcal{M}}} |d_2^T u(o)|$ **then**
30: $d_2 \leftarrow s$
31: **end if**
32: **end if**
33: **end for**
34: $R \leftarrow [d_1, d_2, d_1 \times d_2]^T$
35: **return** (R, -Rp)

B.4 Lattice Centering

By our definition above, the lattice defined by a coordinate frame places corners of cubes at points all of whose coordinates are integer multiples of r. Given a particular conformation, it may be better to shift the lattice by a length of $r/2$ in a particular direction, recentering the lattice cubes.

Algorithm B.5 CENTERFRAME$(T, \mathcal{M}, c, c_f, c_t)$: recenter the coordinate frame T for the set of atoms $\mathcal{M}$ and conformation c using parameters: centering fraction c_f, resolution r, centering tolerance c_t.

$l \leftarrow \lceil c_f |\mathcal{M}| \rceil$
set x_l to be the lth lowest element in the set $\{|T(c(a))_x|, a \in \mathcal{M}\}$
set y_l to be the lth lowest element in the set $\{|T(c(a))_y|, a \in \mathcal{M}\}$
set z_l to be the lth lowest element in the set $\{|T(c(a))_z|, a \in \mathcal{M}\}$
$\mathcal{A} \leftarrow \{a \in \mathcal{M}$ such that $|T(c(a))_x| \leq x_l, |T(c(a))_y| \leq y_l, |T(c(a))_z| \leq z_l\}$
define

$$
\begin{aligned}
v_1 &\leftarrow (r/2, \; r/2, \; r/2) \\
v_2 &\leftarrow (0, \; r/2, \; r/2) \\
v_3 &\leftarrow (r/2, \; 0, \; r/2) \\
v_4 &\leftarrow (0, \; 0, \; r/2) \\
v_5 &\leftarrow (r/2, \; r/2, \; 0) \\
v_6 &\leftarrow (0, \; r/2, \; 0) \\
v_7 &\leftarrow (r/2, \; 0, \; 0) \\
v_8 &\leftarrow (0, \; 0, \; 0)
\end{aligned}
$$

for $i \leftarrow 1$ **to** 8 **do**
  define the coordinate frame $T_i(p) = p + v_i$
  $Q_i \leftarrow$ CUBIFY$(\mathcal{M}, c, T_i \circ T, r, c_t)$
  $n_i \leftarrow |Q_i|$
  $d_{x,i} \leftarrow \max_{q \in Q_i} q_x - \min_{q \in Q_i} q_x$
  $d_{y,i} \leftarrow \max_{q \in Q_i} q_y - \min_{q \in Q_i} q_y$
  $d_{z,i} \leftarrow \max_{q \in Q_i} q_z - \min_{q \in Q_i} q_z$
end for
define j to be the index of the least member of the set $\{(n_i, d_{z,i}, d_{y,i}, d_{x,i}\}$, where comparisons are done using a dictionary ordering (that is, compare the first component, if case of equality compare the second component, etc.)
return (T_j)

B.5 Lattice Perturbations

The base coordinate frame is not necessarily optimal for quantization, so a set of "close" frames are also examined. By parameterizing the space of coordinate frames $SO_3(\mathbb{R}) \times \mathbb{R}^3$ into 3 rotational dimensions and 3 translational dimensions and taking n_r equal steps in each rotational dimension and n_t equal steps in each translational dimension, a set of $n_r^3 n_t^3$ perturbation frames is built. These can later be composed with the base coordinate frame to give the desired set.

Algorithm B.6 PERTURBATIONFRAMES(n_t, v_t, n_r, v_r): generate a set of perturbation coordinates frames using parameters: resolution r, number of translations n_t, translational variance v_t, number of rotations n_r, rotational variance v_r.

$\mathcal{T} \leftarrow \emptyset$
for $i_x \leftarrow 0$ **to** $n_r - 1$ **do**
  define R_x to be the coordinate frame corresponding to rotation about the x-axis by $\pi v_r(2i_x + 1 - n_r)/(4n_r)$ radians
  for $i_y \leftarrow 0$ **to** $n_r - 1$ **do**
    define R_y to be the coordinate frame corresponding to rotation about the y-axis by $\pi v_r(2i_y + 1 - n_r)/(4n_r)$ radians
    for $i_z \leftarrow 0$ **to** $n_r - 1$ **do**
      define R_z to be the coordinate frame corresponding to rotation about the z-axis by $\pi v_r(2i_z + 1 - n_r)/(4n_r)$ radians
      for $j_x \leftarrow 0$ **to** $n_t - 1$ **do**
        define the coordinate frame

$$T_x(p) = p + (rv_t(2j_x + 1 - n_r)/(2n_r), 0, 0)$$

        for $j_y \leftarrow 0$ **to** $n_t - 1$ **do**
          define the coordinate frame

$$T_y(p) = p + (0, rv_t(2j_y + 1 - n_r)/(2n_r), 0)$$

          for $j_z \leftarrow 0$ **to** $n_t - 1$ **do**
            define the coordinate frame

$$T_y(p) = p + (0, 0, rv_t(2j_z + 1 - n_r)/(2n_r))$$

            add $T_z \circ T_y \circ T_x \circ R_z \circ R_y \circ R_x$ to $\mathcal{T}$
          end for
        end for
      end for
    end for
  end for
end for
return $(\mathcal{T})$

B.6 Cubification

Given a conformation and a lattice (defined by a coordinate frame and resolution), cubification is the process of determining which lattice cubes are filled by the conformation. The set of lattice cubes is constructed by taking any cube in which an atom center in the conformation directly falls and also cubes which are sufficiently close to the van Der Waals sphere of an atom.

Algorithm B.7 CUBIFY($\mathcal{M}, c, T, r, t$): quantize the conformation c of the molecular structure $\mathcal{M}$ into a set of cubes defined on the coordinate frame T using parameters: coordinate frame T, resolution r, tolerance t.

$\mathcal{Q} \leftarrow \emptyset$ {$\mathcal{Q}$ will contain the set of cubes in the quantization. }
define a function $d : \mathcal{M} \rightarrow \mathbb{R}$ that takes each atom to its van Der Waals radius
for all $a \in \mathcal{M}$ **do**
  $p \leftarrow T(c(a))$ {the transformed position }
  $q \leftarrow (\lfloor p_x/r \rfloor, \lfloor p_y/r \rfloor, \lfloor p_z/r \rfloor)$ {the integer coordinates of the cube into which the atom falls }
  if $q \notin \mathcal{Q}$ **then**
    add q to $\mathcal{Q}$
  end if
  define $\tilde{\mathcal{Q}}$ to be the set of cubes neighboring q:

$$\tilde{\mathcal{Q}} \leftarrow \{\hat{q} \in \mathbb{Z}^3 \text{ such that } \max(|q_x - \hat{q}_x|, |q_y - \hat{q}_y|, |q_z - \hat{q}_z|) = 1\}$$

  for all $\hat{q} \in \tilde{\mathcal{Q}}$ **do**
    define $\hat{d}$ to be the distance of the point in $\mathbb{R}^3$ in the cube defined by $\hat{q}$ that is closest to p:

$$\hat{d} \leftarrow \min_{v \in [rq_x, rq_x+r) \times [rq_y, rq_y+r) \times [rq_z, rq_z+r)} \|p - v\|$$

    if $\hat{d} - d(a) < tr$ **and** $\hat{q} \notin \mathcal{Q}$ **then**
      add $\hat{q}$ to $\mathcal{Q}$ {add a neighboring cube if it is too close to the van Der Waals sphere of the atom }
    end if
  end for
end for
return ($\mathcal{Q}$)

APPENDIX B. QUANTIZATION

B.7 Entire Process

Given a set of conformations, in QUANTIZE we see the entire quantization algorithm. For each conformation:

1. a base frame and centering frame are calculated
2. perturbations of the base frame are used in order to find the lattice that results first in the smallest number of cubes and second with the least distance from the base frame
3. using the atomic functionality map, functionalities are assigned to the cubes in the minimal quantization

APPENDIX B. QUANTIZATION

Algorithm B.8 QUANTIZE($\mathcal{M}, \mathcal{C}, r, t, r_f, r_m, q_r, c_f, c_t, n_t, v_t, n_r, v_r$): quantize the set of conformations $\mathcal{C}$ for the molecular structure $\mathcal{M}$ with parameters: resolution r, tolerance t, ring factor r_f, ring minimum r_m, radius factor q_r, centering fraction c_f, centering tolerance c_t, number of translations n_t, translational variance v_t, number of rotations n_r, rotational variance v_r, polarizable minimum p_m.

1: $(f, \mathcal{E}_q) \leftarrow$ ATOMFUNCTIONMAP($\mathcal{M}$) {build a functionality map and a list of atoms to be excluded from quantization }
2: $\mathcal{S} \leftarrow \emptyset$ {will hold the resulting quantizations }
3: **for all** $c \in \mathcal{C}$ **do**
4: $\quad \tilde{\mathcal{M}} \leftarrow$ FRAMEATOMS($\mathcal{M} - \mathcal{E}_q, c, r_f, r_m, q_r$)
5: $\quad T_b \leftarrow$ BASEFRAME($\tilde{\mathcal{M}}, c$) {base coordinate frame }
6: $\quad T_c \leftarrow$ CENTERFRAME($T, \mathcal{M} - \mathcal{E}_q, c, c_f, c_t$) {lattice centering }
7: $\quad q_m \leftarrow \infty$ {smallest number of cubes seen so far }
8: $\quad d_m \leftarrow \infty$ {smallest transform distance seen so far }
9: $\quad \mathcal{T} \leftarrow$ PERTURBATIONFRAMES(n_t, v_t, n_r, v_r) {set of perturbation frames }
10: $\quad$ **for all** $T_p \in \mathcal{T}$ **do**
11: $\quad\quad T \leftarrow T_c \circ T_p \circ T_b$ {the total coordinate transform }
12: $\quad\quad \mathcal{Q} \leftarrow$ CUBIFY($\mathcal{M} - \mathcal{E}_q, c, T, r, t$) {cubes given this transform }
13: $\quad\quad q \leftarrow |\mathcal{Q}|$ {number of cubes }
14: $\quad\quad d \leftarrow \sum_{a \in \mathcal{M} - \mathcal{E}_q} \|T_p(c(a)) - c(a)\|$ {transform distance }
15: $\quad\quad$ **if** $|\mathcal{Q}| < q_m$ **or** $(|\mathcal{Q}| = q_m$ **and** $d < d_m)$ **then**
16: $\quad\quad\quad q_m \leftarrow q, d_m \leftarrow d, \mathcal{Q}_m \leftarrow \mathcal{Q}, T_m \leftarrow T$
17: $\quad\quad$ **end if**
18: $\quad$ **end for**
19: $\quad$ define $f_m : \mathcal{Q}_m \rightarrow \mathcal{F}$ as $f_m(q) = \mathcal{F}_7$.
20: $\quad$ **for all** $a \in \mathcal{M} - \mathcal{E}_q$ **do**
21: $\quad\quad q \leftarrow (\lfloor T_m(c(a))_x / r \rfloor, \lfloor T_m(c(a))_y / r \rfloor, \lfloor T_m(c(a))_z / r \rfloor)$
$\quad\quad$ {q is the the cube a falls into}
22: $\quad\quad$ **if** $f(a)$ has higher priority than $f_m(q)$ **then**
23: $\quad\quad\quad$ **if** $f(a) \neq \mathcal{F}_6$ **then**
24: $\quad\quad\quad\quad f_m(q) \leftarrow f(a)$
25: $\quad\quad\quad$ **else if** at least p_m atoms with $\mathcal{F}_6$ functionality are in q **then**
26: $\quad\quad\quad\quad f_m(q) \leftarrow f(a)$
27: $\quad\quad\quad$ **end if**
28: $\quad\quad$ **end if**
29: $\quad$ **end for**
30: $\quad$ add $(\mathcal{Q}_m, f_m)$ to $\mathcal{S}$
31: **end for**
32: **return** $(\mathcal{S})$

Appendix C

Surface Complementarity

In order to view molecules in the space of theoretical surfaces, we must establish complementarity between quantized conformations and theoretical surfaces. This is done in FITSURFACES as follows:

1. all 24 possible orientations of each quantizaed conformation are considered
2. for each orientation, the quantized conformation is shifted down and below the plane
3. a set of surfaces which fit each shifted conformation are detected
4. functionalities for each surface are computed such that the binding energy between the surface and the quantized conformation are favorable

APPENDIX C. SURFACE COMPLEMENTARITY

Algorithm C.1 FITSURFACES($\mathcal{Q}, f_q, E_c, r_b, \ldots$): calculate all surfaces with functionality that are complementary to the quantizated conformation $\mathcal{Q}$ with functionality map f_q, conformational energy E_c, and r_b rotatable bonds using the following parameters: minimum surface opening area A, maximum surface volume V, area-threshold A_t, max-non-central M_{nc}, max-contiguous M_c, max-extrustion M_e, number of points of characteristic functionality n_f, minimum energy E_{min}, minimum fit quanta q_{min}, minimum slackness s_{min}, maximum slackness s_{max}, maximum protrusion levels p_{max}, translational-rotational-vibrational entropy E_{trv}, rotatable bond coefficient c_r, hydrophobic energy coefficient c_h, hydrophobic surface energy coefficient c_s, potential function P.

1: $\mathcal{S} \leftarrow \emptyset$ {will hold the resulting surfaces }
2: define $\mathcal{R}$ to be the set of 24 lattice rotations
3: **for all** $R \in \mathcal{R}$ **do**
4: $\hat{\mathcal{Q}} \leftarrow \{Rq, q \in \mathcal{Q}\}$ {rotate the quantization by R}
5: $t_x \leftarrow \lfloor(\max_{q \in \hat{\mathcal{Q}}} q_x + \min_{q \in \hat{\mathcal{Q}}} q_x)/2\rfloor$
6: $t_y \leftarrow \lfloor(\max_{q \in \hat{\mathcal{Q}}} q_y + \min_{q \in \hat{\mathcal{Q}}} q_y)/2\rfloor$
7: $t_z \leftarrow \min_{q \in \hat{\mathcal{Q}}} q_z$
8: $\hat{\mathcal{Q}} \leftarrow \{(q_x - t_x, q_y - t_y, q_z - t_z), q \in \hat{\mathcal{Q}}\}$ {recenter over the x-y plane}
9: **for all** $d \leftarrow 0$ **to** $\max_{q \in \hat{\mathcal{Q}}} q_z$ **do**
10: $\hat{\mathcal{Q}}_d \leftarrow \{(q_x, q_y, q_z - d), q \in \hat{\mathcal{Q}}$ such that $q_z \leq d\}$ {shift the quantization down and only keep cubes on or under the x-y plane}
11: **if** $|\hat{\mathcal{Q}}_d| < \min(q_{min}, |\mathcal{Q}|) - s_{min}$ **then** {skip if not enough cubes are below the plane}
12: **goto step** 9
13: **end if**
14: **if** $\max_{q \in \hat{\mathcal{Q}}_d} q_z > p_{max}$ **then** {skip if too many layers of cubes are sticking out of the plane}
15: **goto step** 9
16: **end if**
17: $S_d \leftarrow$ CORESURFACE($\hat{\mathcal{Q}}_d$)
18: $A_d \leftarrow \min(s_{max} + |S_d|, A), V_d \leftarrow \min(s_{max} + |S_d|, V)$
19: **for all** $S \in$ DETECTSURFACES($S_d, A_d, V_d, A_t, M_{nc}, M_c, M_e$) **do**
20: **for all** $(S_f, f_s) \in$ FUNCTIONALIZESURFACE(S, n_f) **do**
21: **if** ENERGY($S_f, f_s, \hat{\mathcal{Q}}_q, f_q, E_c, r_b, r, E_{trv}, c_r, c_h, c_s, P$) $\geq E_{min}$ **then**
22: **if** no translation or rotation of (S_f, f_s) is in $\mathcal{S}$ **then**
23: add (S_f, f_s) to $\mathcal{S}$
24: **end if**
25: **end if**
26: **end for**
27: **end for**
28: **end for**
29: **end for**
30: **return** ($\mathcal{S}$)

APPENDIX C. SURFACE COMPLEMENTARITY

Algorithm C.2 CORESURFACE($\mathcal{Q}$): calculate the minimal surface fitting a quantized conformation.

```
1: O ← ∅{surface opening}
2: for all q ∈ Q do
3:    add (q_x, q_y) to O, if not already present
4: end for
5: define a depth function d : O → ℕ such that d(x, y) = 1
6: for all q ∈ Q do
7:    d(q_x, q_y) ← max(d(q_x, q_y), 1 − q_z)
8: end for
9: return (shape(O, d))
```

APPENDIX C. SURFACE COMPLEMENTARITY

Algorithm C.3 DETECTSURFACES($S_c, A, V, A_t, M_{nc}, M_c, M_e$): detect additional surfaces by adding cubes to S_c subject to parameters: minimum surface opening area A, maximum surface volume V, area-threshold A_t, max-non-central M_{nc}, max-contiguous M_c, max-extrustion M_e.

1: $\mathcal{S} \leftarrow \emptyset$ {the detected surfaces}
2: $\mathcal{S}_n \leftarrow \{S_c\}$
3: **while** $\mathcal{S}_n \neq \emptyset$ **do**
4: $S_m \leftarrow \text{pop}(\mathcal{S}_n), O \leftarrow \text{opening}(S_m), d_m \leftarrow \text{depth}(S)$
5: **if** O is connected **and** $\text{area}(O) \leq A$ **then** {add surfaces obtained by adding cubes below S_m}
6: $d \leftarrow d_m, v \leftarrow \text{volume}(S_m)$
7: **while** $v \leq V$ **do**
8: $S \leftarrow \text{shape}(O, d)$
9: add S to $\mathcal{S}$
10: **for** $x \leftarrow -\text{area}(O)$ **to** $\text{area}(O), y \leftarrow -\text{area}(O)$ **to** $\text{area}(O)$ **do**
11: **if** $(x, y) \in O$ **and** $v + 1 \leq V$ **then**
12: $v \leftarrow v + 1, d(x, y) \leftarrow d(x, y) + 1$
13: **goto step** 7
14: **else**
15: $v \leftarrow v - d(x, y) + d_m(x, y), d(x, y) \leftarrow d_m(x, y)$
16: **end if**
17: **end for**
18: **goto step** 20{no more surfaces left}
19: **end while**
20: **end if**
21: **if** $\text{area}(O) < A$ **and** $\text{volume}(S_m) < V$ **then** {consider surfaces made by enlarging the opening}
22: define $\mathcal{P}$ to be the set of all possible openings obtained by adding a single square to O adjacent to a square already present in O
23: **for all** $P \in \mathcal{P}$ **do**
24: define $d_p : P \rightarrow \mathbb{N}$ such that $d_p(x, y) = d_m(x, y)$ for $(x, y) \in O$ and $d_p(x, y) = 1$ otherwise
25: add $\text{shape}(P, d_p)$ to $\mathcal{S}_n$
26: **end for**
27: **end if**
28: **end while**
29: $\mathcal{O} \leftarrow$ OPENINGFILTER($\{\text{opening}(S), S \in \mathcal{S}\}, A_t, M_{nc}, M_c$)
30: **for all** $S \in \mathcal{S}$ **do** {filter surfaces based on openings}
31: **if** $\text{opening}(S) \notin \mathcal{O}$ **then**
32: delete S from $\mathcal{S}$
33: **end if**
34: **end for**
35: $\mathcal{S} \leftarrow$ SHAPEFILTER($\mathcal{S}, M_e$){filter surfaces based on shape}
36: **return** ($\mathcal{S}$)

APPENDIX C. SURFACE COMPLEMENTARITY

C.1 Energy

Complementarity energy between a quantization of a molecular conformation and a a cubic theoretical surface is the sum of several components:

- translational-vibration-rotational entropy (a constant)
- a conformational energy term, representing the energy of this conformation relative to the minimal energy conformation (calculated by the conformational generator)
- a term proportional to the number of rotatable bonds
- potential energy, the sum of energy due to interactions between the functionalities of overlapping negative space surface cubes and positive space quantization cubes (represented as a function $P : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R} \cup \{-\infty\}$)
- hydrophobic energy due to the exclusion of water from cubes in the surface, proportional to the surface area from which water is excluded

APPENDIX C. SURFACE COMPLEMENTARITY

Algorithm C.4 ENERGY$(S, f_s, \mathcal{Q}, f_q, E_c, r_b, r, E_{trv}, c_r, c_h, c_s, P)$: calculate the complementarity energy between a surface S with functionality map f_s and a quantized conformation $\mathcal{Q}$ with functionality map f_q, conformational energy E_c, and r_b rotatable bonds using parameters: resolution r, translational-rotational-vibrational entropy E_{trv}, rotatable bond coefficient c_r, hydrophobic energy coefficient c_h, hydrophobic surface energy coefficient c_s, potential function P.

$E_r \leftarrow c_r r_b$ {energy due to rotatable bonds}
$E_p \leftarrow 0$ {energy due to potential interactions}
for all $s \in S$ **do**
  if $s \in \mathcal{Q}$ **then**
    $E_p \leftarrow E_p + P(f_q(s), f_s(s))$
  end if
end for
$E_h \leftarrow 0$ {energy due to hydrophobicity}
for all $s \in S \cap \mathcal{Q}$ **do**
  define

$$\mathcal{T} \leftarrow \{ \begin{array}{l} (s_x+1, s_y, s_z), (s_x-1, s_y, s_z), \\ (s_x, s_y+1, s_z), (s_x, s_y-1, s_z), \\ (s_x, s_y, s_z-1) \end{array} \}$$

  $a \leftarrow r^2 |\mathcal{T} \cap \overline{S}|$ {the area of contact between the surface and the quantized conformation at this point}
  if $f_s(s) = \mathcal{F}_8$ **then** {hydrophobic energy due to slightly hydrophobic surface cubes}
    $E_h \leftarrow E_h + c_h c_s a$
  else if $f_s(s) \in \{\mathcal{F}_6, \mathcal{F}_7\}$ **then** {hydrophobic energy due to non-polar surface cubes}
    $E_h \leftarrow E_h + c_h a$
  end if
  if $f_q(s) \in \{\mathcal{F}_6, \mathcal{F}_7\}$ **then** {hydrophobic energy due to non-polar quantized conformation cubes}
    $E_h \leftarrow E_h + c_h a$
  end if
end for
return $(E_{trv} + E_r + E_p + E_h - E_c)$

Appendix D

Molecular Library Comparison

A library is a set of molecular structures. Given a library, the set of complementary theoretical surfaces is defined as the union of all surface shape/functionality pairs complementary to any quantized conformation of any molecule in the library.

The algorithm LIBRARYCOMPARE calculates a score proportional to the similarity of the two libraries. The score is calculated by representing each library as its set of complementary theoretical surfaces, and using the SIMILARITYSCORE primitive to determine the similarity or dissimilarity two sets of theoretical surfaces. If the molecular libraries each contain only one molecule, then the algorithm calculates a score proprotional to the similarity of the two molecules.

Algorithm D.1 LIBRARYCOMPARE($\mathcal{L}_1, \mathcal{L}_2$): calculate a similarity score between 0 and 1000 for two molecular libraries.

1: define $\mathcal{T}_1$ to be the theoretical surfaces complementary to $\mathcal{L}_1$
2: define $\mathcal{T}_2$ to be the theoretical surfaces complementary to $\mathcal{L}_2$
3: **return** (SIMILARITYSCORE($\mathcal{T}_1, \mathcal{T}_2$))

APPENDIX D. MOLECULAR LIBRARY COMPARISON

Algorithm D.2 SIMILIARITYSCORE($\mathcal{T}_1, \mathcal{T}_2$): calculate a similarity score between 0 and 1000 for two sets of theoretical surfaces.

1: define sets of complementary surface shapes:

$$\begin{aligned} \mathcal{S}_1 &\leftarrow \{S \text{ such that for some } f, (S, f) \in \mathcal{T}_1\} \\ \mathcal{S}_2 &\leftarrow \{S \text{ such that for some } f, (S, f) \in \mathcal{T}_2\} \end{aligned}$$

For purposes of comparing two theoretical surfaces or surface shapes, note that translations and x-y plane rotations of a surface are considered identical to the original surface.

2: $s_s \leftarrow |\mathcal{S}_1 \cap \mathcal{S}_2| / |\mathcal{S}_1 \cup \mathcal{S}_2|$ {shape score}
3: $s_f \leftarrow 0$ {functionality score}
4: **for all** $S \in \mathcal{S}_1 \cap \mathcal{S}_2$ **do**
5: define complementary functionalities for this surface shape:

$$\begin{aligned} F_1 &\leftarrow \{f \text{ such that } (S, f) \in \mathcal{T}_1\} \\ F_2 &\leftarrow \{f \text{ such that } (S, f) \in \mathcal{T}_2\} \end{aligned}$$

6: define sets of "active" cubes, that is cubes with non-default functionality:

$$\begin{aligned} \mathcal{Q}_1 &\leftarrow \{Q \subset \mathbb{Z}^3 \text{ such that } \exists f \in F_1, \forall q \in Q, f(q) \neq \mathcal{F}_8\} \\ \mathcal{Q}_2 &\leftarrow \{Q \subset \mathbb{Z}^3 \text{ such that } \exists f \in F_2, \forall q \in Q, f(q) \neq \mathcal{F}_8\} \\ \mathcal{Q}_i &\leftarrow \{Q \subset \mathbb{Z}^3 \text{ such that } \exists f \in F_1 \cap F_2, \forall q \in Q, f(q) \neq \mathcal{F}_8\} \end{aligned}$$

7: $s_f \leftarrow s_f + |\mathcal{Q}_i| / (|\mathcal{S}_1 \cap \mathcal{S}_2| \min(|\mathcal{Q}_1|, |\mathcal{Q}_2|))$
8: **end for**
9: **return** $(1000 s_s s_f)$

Appendix E

Protein Quantization

In a process complementary to the quantization of a molecular conformation, the target sites of a protein surface are quantized into the same negative space cubic representation used by theoretical surfaces. This allows the following analyses:

- comparison between a set of known protein target sites and the set of all possible theoretical surfaces within given parameters of volume, shape, and functionality
- comparison between two different sets of known protein target sites.
- comparison between a set of known protein target sites and a set of theoretical surfaces to which a given set of molecules is complementary

The protein quantization process is accomplished in the following steps, as depicted in Figure 30:

1. A 3D crystal structure of the protein is examined and a functionality map is built for the protein atoms using the algorithm ATOMFUNCTIONALMAP.
2. A protein surface is generated from the 3D structure. A protein surface is a set of triangles defining the surface of the protein that is accessible to water molecules (known as the Connolly surface). Michael Connolly's MSRoll software is an example of a package that can generate a protein surface suitable for this purpose.
3. Subsets of the surface which are target sites likely for the binding of small molecules are detected. This can be accomplished, for example, by looking for highly concave regions. Michael Connolly's MSForm software is an example of a package that can measure surface curvature and detect pockets suitable for this purpose.
4. Each target site is isolated and examined individually.
5. Each target site is quantized into a set of negative space cubes with associated functionalities using the protein function map and the algorithm TARGETSITE-QUANTIZE. The underingly process is very similar to the algorithm QUANTIZE,

using an imaginary molecule that is defined by atoms centered at a set of points with a given radius that fill the target site.

6. Each set of quantized negative space cubes with functionality is converted to a set of theoretical surfaces satisfying the proper constraints (for example, no occluded cubes are allowed) using the algorithm BUILDSURFACES.

Algorithm E.1 TARGETSITEQUANTIZE($\mathcal{T}, f, v, r, r_l, r_v, b, n_p, s_r, \ldots$): calculate a negative space cubic representation of the target site defined by the triangles in set $\mathcal{T}$, with v as a normal vector pointing out of the target site, f as a functionality map for the entire protein, using parameters: resolution r, lattice density r_l, lattice van Der Waals radius r_v, lattice tolerance t_l, number of lattice neighbors n_p, buffer distance b, search radius s_r, and additional parameters for subrountine calls (see below) as necessary.

---

define a coordinate frame T_l with origin at the center of the target site, z axis in the direction of v, x and y axes determined by the longest side of the pocket
define a set of points $\mathcal{P} \subset \mathbb{R}^3$ to be the points on a lattice with coordinate frame T_l and cube side length r_l such that $p \in \mathcal{P}$ iff p is contained in the target site and the closest triangle in $\mathcal{T}$ is at least distance b away from p
remove from $\mathcal{P}$ all points who do not have at least n_p neighboring points also in $\mathcal{P}$, where each point has neighbors consisting of the 26 points on the lattice offset by at most one cube from the point in question
if $\mathcal{P}$ is disconnected, consider each connected component seperately in the following steps
define a "molecular" structure $\mathcal{M}$ with conformation c by imagining $\mathcal{P}$ to be a set of atoms with van Der Waals radius r_v
define a base coordinate frame T_b using the algorithm BASEFRAME on $\mathcal{M}$ and c
define a centering coordinate frame T_c using the algorithm CENTERFRAME on $\mathcal{M}$ and c
define a set of perturbation frames using PERTURBATIONFRAMES
as in QUANTIZE, by examining all perturbation frames, find the total frame which minimizes the number of negative space cubes of resolution r (calculated using CUBIFY with tolerance t_l) and the total transformation distance
denote the above set of negative space cubes by $\mathcal{Q}$
redefine the coordinate system for the negative space cubes in $\mathcal{Q}$, such that the z axis is the direction with the greatest component in the v direction, and the maximum z value is 0
build a functionality map $f_q : \mathcal{Q} \rightarrow \mathcal{F}$ by, for each $q \in \mathcal{Q}$, finding the highest priority functionality associated by f with an atom in the protein within search radius s_r to the center of q, or assigning $\mathcal{F}_8$ if there are no such atoms
return $(\mathcal{Q}, f_q)$

APPENDIX E. PROTEIN QUANTIZATION

Algorithm E.2 BUILDSURFACES($\mathcal{Q}, f_q, m_o, n_f \ldots$): return the set of theoretical surfaces satifying proper constraints described by the negative space cubes $\mathcal{Q}$ with functionality f_q, using parameters: maximum occlusions m_o, number of functionality points n_f, additional parameters for subroutine calls (see below) as necessary.

define $O \leftarrow \emptyset$ and $d : \mathbb{Z}^2 \rightarrow \mathbb{Z}$ such that $d(x, y) = 0$ {future surface opening and depths}
for all $q \in \mathcal{Q}$ **do** {build an opening and depth function}
  if $(q_x, q_y) \notin O$ **then**
    add (q_x, q_y) to O, $d(q_x, q_y) \leftarrow \max(d(q_x, q_y), 1 - q_z)$
  end if
end for
$o_c \leftarrow$ REMOVEOCCLUDEDCUBES($\mathcal{Q}, O, d$)
if $o_c > m_o$ **then**
  return ($\emptyset$){too many occluded cubes}
end if
if O is disconnected **then**
  define $\mathcal{O}$ to be the set of connected components, $\mathcal{S} \leftarrow \emptyset$
  for all $O_c \in \mathcal{O}$ **do** {generate a set of surfaces for each connected component}
    $\mathcal{Q}_c \leftarrow \{q \in \mathcal{Q}$ such that $(q_x, q_y) \in O_c\}$
    $\mathcal{S} \leftarrow \mathcal{S} \cup$ BUILDSURFACES($\mathcal{Q}_c, f_q$)
  end for
  return ($\mathcal{S}$)
end if
if O does not satisfy filtering rules imposed by OPENINGFILTER **then**
  return ($\emptyset$)
end if
$S \leftarrow$ shape(O, d)
if S does not satisfy filtering rules imposed by SHAPEFILTER **then**
  return ($\emptyset$)
end if
$\mathcal{S} \leftarrow \emptyset$
$\mathcal{Q}_a \leftarrow \{q \in \mathcal{Q}$ such that $f_q(q) \neq \mathcal{F}_8\}$ {set of negative space cubes with non-default functionality}
for all $Q \subset \mathcal{Q}_a$ such that $|Q| = n_f$ **do**
  define $f_s : S \rightarrow \mathcal{F}$ such that $f_s(x, y, z) = f_q(x, y, z)$ for $(x, y) \in \mathcal{Q}_a$, $f_s(x, y, z) = \mathcal{F}_8$ otherwise, add (S, f_s) to $\mathcal{S}$
end for
return ($\mathcal{S}$)

APPENDIX E. PROTEIN QUANTIZATION

Algorithm E.3 REMOVEOCCLUDEDCUBES($\mathcal{Q}, O, d$): remove from $\mathcal{Q}$ cubes which are occluded, adjusting opening O and depths d, return the number of occluded cubes removed.

```
1: o_c ← 0{count of occluded cubes}
2: for all (x, y) ∈ O do
3:    for z ← −1 to 1 − d(x, y) do
4:       if (x, y, z) ∈ Q and (x, y, z + 1) ∉ Q then
5:          o_c ← o_c + 1
6:          remove (x, y, z) from Q
7:       end if
8:    end for
9:    Z_xy ← {q_z, (x, y, q_z) ∈ Q}
10:   if Z_xy = ∅ then
11:      delete (x, y) from O, d(x, y) ← 0
12:   else
13:      d(x, y) ← min{1 − q_z, q_z ∈ Z_xy}
14:   end if
15: end for
16: return (o_c)
```

Appendix F

Protein Surface Comparison

A target surface set is defined as the set of all theoretical target surfaces to which a set of known protein surfaces map. The target surface set may comprise, for example, all of the surfaces mapped from one protein, all of the surfaces mapped from multiple proteins, or all of the surfaces mapped from specific sites on multiple proteins.

The algorithm PROTEINCOMPARE calculates a score proportional to the similarity of the two sets of protein surfaces. The score is calculated by representing each protein surface set as its target surface set, and using the SIMILARITYSCORE primitive to determine the similarity or dissimilarity two sets of theoretical surfaces.

Algorithm F.1 PROTEINCOMPARE($\mathcal{P}_1, \mathcal{P}_2$): calculate a similarity score between 0 and 1000 for two protein surface sets.

1: define $\mathcal{T}_1$ to be the target surface set corresponding to $\mathcal{P}_1$
2: define $\mathcal{T}_2$ to be the target surface set corresponding to $\mathcal{P}_2$
3: **return** (SIMILARITYSCORE($\mathcal{T}_1, \mathcal{T}_2$))

Appendix G

Protein/Library Comparison

The algorithm PROTEINLIBRARYCOMPARE calculates a score proportional to the complementarity of a library of small molecules and a set of protein surfaces. The score is calculated by representing the protein surface set as the theoretical surface set to which it is similar, the molecular library as the theoretical surface set to which it is complementary, and using the SIMILARITYSCORE primitive to determine the similarity or dissimilarity two sets of theoretical surfaces.

Algorithm G.1 PROTEINLIBRARYCOMPARE($\mathcal{L}, \mathcal{P}$): calculate a complementarity score between 0 and 1000 for a molecular library $\mathcal{L}$ and a protein surface set $\mathcal{P}$.

1: define $\mathcal{T}_l$ to be the theoretical surfaces complementary to $\mathcal{L}$
2: define $\mathcal{T}_p$ to be the target surface set corresponding to $\mathcal{P}$
3: **return** (SIMILARITYSCORE($\mathcal{T}_l, \mathcal{T}_p$))

Appendix H

Parameter Values

The values of parameters used are as follows:

- Maximum opening area (A): 15
- Maximum surface volume (V): 18
- Area threshold: (A_t): 8
- Maximum non-central opening squares (M_{nc}): 5
- Maximum contiguous opening squares (M_c): 3
- Maximum surface extrusions (M_e): 3
- Number of surface cubes of specific functionality (n_f): 4
- Maximum number of conformations per molecule: 300
- Resolution (r): 4.24 Angstroms
- Tolerance (t): 0.32
- Ring factor (r_f): 0
- Ring minimum (r_m): 13
- Radius factor (q_r): 0.75
- Centering tolerance (c_t): 0.1
- Centering fraction (c_f): 0.75
- Number of translations (n_t): 5
- Translational variance (v_t): 0.2
- Number of rotations (n_r): 5

APPENDIX H. PARAMETER VALUES

- Rotational variance (v_r): 0.1
- Polarizable minimum (p_m): 2
- Minimum fit quanta (q_{min}): 9
- Minimum slackness (s_{min}): 2
- Maximum slackness (s_{max}): 0
- Maximum protrusion levels (p_{max}): 1
- Minimum energy (E_{min}): 8.0 kCal
- Translational-rotation-vibrational entropy (E_{trv}): -9.0 kCal
- Rotatable bond coefficient (r_c): -0.7 kCal/bond
- Hydrophobic energy coefficient (c_h): 0.025 kCal/Angstrom^2
- Hydrophobic surface energy coefficient (c_s): 0.8
- Potential function (P): kCal

Molecule \ Surface	$\mathcal{F}_1$	$\mathcal{F}_2$	$\mathcal{F}_3$	$\mathcal{F}_4$	$\mathcal{F}_5$	$\mathcal{F}_6$	$\mathcal{F}_7$	$\mathcal{F}_8$
$\mathcal{F}_1$	$-\infty$	4.0	$-\infty$	2.0	$-\infty$	$-\infty$	$-\infty$	$-\infty$
$\mathcal{F}_2$	4.0	$-\infty$	$-\infty$	$-\infty$	2.0	0.0	$-\infty$	$-\infty$
$\mathcal{F}_3$	0.0	0.0	$-\infty$	2.5	2.5	$-\infty$	$-\infty$	0.0
$\mathcal{F}_4$	2.0	$-\infty$	$-\infty$	$-\infty$	3.5	$-\infty$	$-\infty$	-1.0
$\mathcal{F}_5$	$-\infty$	2.0	$-\infty$	3.5	$-\infty$	$-\infty$	$-\infty$	-0.5
$\mathcal{F}_6$	$-\infty$	0.0	$-\infty$	$-\infty$	$-\infty$	2.5	0.0	0.3
$\mathcal{F}_7$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	0.0	0.5	0.2

- Lattice density (r_l): 1.5 Angstroms
- Lattice van Der Waals radius (r_v): 0.75 Angstroms
- Lattice tolerance (t_l): 0.2
- Buffer distance (b): 0.5 Angstroms
- Number of lattice neighbors (n_p): 3
- Search radius (s_r): 2.22 Angstroms
- Maximum number of occluded cubes (m_o): 1

CLAIMS

1. A computer-based method comprising

defining a set of constraints on possible target surfaces,

defining a fully enumerated set of theoretical target surfaces under the defined constraints, such that each surface has a defined, continuous volume and a defined, continuous surface area,

mapping one or more sets of objects to the fully enumerated set of theoretical target surfaces to define corresponding subsets of the fully enumerated set of theoretical target surfaces, and

analyzing an aspect of diversity of the objects based on degrees of similarities and differences among the corresponding subsets.

2. The method of claim 1 in which the target surfaces comprise negative space target surfaces.

3. The method of claim 1 in which the objects comprise positive space object surfaces associated with different molecules.

4. The method of claim 2 in which the objects comprise positive space object surfaces associated with different molecules and in which the objects are mapped by

defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which positive space object surfaces of conformations of molecules are complementary, and

the aspect of diversity that is analyzed is the difference or similarity between the molecules which map to those negative space theoretical target surfaces.

5. The method of claim 1 in which the objects comprise negative space object surfaces associated with different proteins.

6. The method of claim 2 in which the objects comprise negative space object surfaces associated with different proteins and in which the objects are mapped by

defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which negative space object surfaces of protein pockets are similar, and the aspect of diversity that is analyzed is the difference or similarity between protein pockets which map to those negative space theoretical target surfaces.

7. The method of claim 1 in which the objects comprise positive space object surfaces associated with different molecules and negative space object surfaces associated with different proteins.

8. The method of claim 2 in which the objects comprise positive space object surfaces associated with different molecules and negative space object surfaces associated with different proteins and in which,

in the case of molecules, the objects are mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which positive space object surfaces of conformations of molecules are complementary,

in the case of proteins, the objects are mapped by defining corresponding subsets of the fully enumerated set of negative space theoretical target surfaces to which negative space object surfaces of protein pockets are similar, and

the aspect of diversity that is analyzed is the difference or similarity of the molecules which map to those negative space theoretical target surfaces to the protein pockets which map to those negative space theoretical target surfaces.

9. The method of claim 1 in which the theoretical target surfaces comprise polyhedrons.

10. The method of claim 1 in which the objects comprise polyhedrons.

11. The method of claim 9 or 10 in which the polyhedrons comprise cubes.

12. The method of claim 9 or 10 in which the polyhedrons are all of the same size and shape.

13. The method of claim 1 in which the set of all theoretical target surfaces defines a diversity space within which the diversity of objects can be measured by mapping those objects to the diversity space.

14. The method of claim 13 also including identifying regions of the diversity space to which no objects map.

15. The method of claim 14 also including designing molecules that occupy at least one of the unfilled theoretical target surfaces of the diversity space.

16. The method of claims 4 or 8 in which complementarity is associated with binding affinities of positive space object surfaces of conformations of molecules to negative space theoretical target surfaces.

17. The method of claim 1 in which the constraints comprise volume.

18. The method of claim 1 in which the constraints comprise associating each of a number of sites of the target surface with a preselected molecular property.

19. The method of claim 18 in which each of the preselected molecular properties is drawn from a larger set of possible molecular properties.

20. The method of claim 18 in which the preselected molecular properties include hydrophobic, polarizable, H-bond acceptor, H-bond donor, H-bond donor/acceptor, potentially positively charged, and potentially negatively charged.

21. The method of claim 18 in which fewer than all of the sites of the target surface are each associated with a different one of the molecular properties and all of the other sites of the target surface are associated with a common molecular property.

22. The method of claim 21 in which the common molecular property comprises slightly hydrophobic.

23. The method of claim 1 in which the degrees of similarities or differences comprise functional properties associated with the corresponding subsets of the fully enumerated set of theoretical target surfaces.

24. The method of claim 1 in which the degrees of similarities or differences comprise shape properties associated with the corresponding subsets of the fully enumerated set of theoretical target surfaces.

25. The method of claim 1 further comprising defining each of the objects by quantizing molecules into polyhedrons.

26. The method of claim 1 also including fitting each of a fixed set of orientations of each conformation of each of the objects to each of the target surfaces.

27. The method of claim 26 further comprising scoring each of the fittings.

28. The method of claim 9 in which the constraints comprise a resolution of the polyhedrons.

29. The method of claim 28 in which the resolution is 4.24 Angstroms.

30. The method of claim 9 in which the constraints comprise maximum and minimum numbers of polyhedrons.

31. The method of claim 9 in which each of the polyhedrons shares a common interface with another of the polyhedrons.

32. The method of claim 1 in which each of the target surfaces has no occlusions of volume greater than a given parameter.

33. The method of claim 1 in which the target surfaces are defined conceptually as having been carved out of a flat surface.

34. A method comprising

categorizing existing molecules based on negative space target surfaces to which conformations of the molecules are complementary, and

designing novel molecules that are complementary to negative space target surfaces to which no conformations of the existing molecular are complementary.

35. A method of creating novel molecules to be tested as ligands for proteins, comprising

categorizing proteins based on target surfaces to which their pockets of known structure map, and

designing novel molecules that are complementary to the negative space target surfaces to which the protein pockets map.

36. A computer programmed to determine the chemical similarity of different molecules, the program comprising

approximating the surface shape of each one of a plurality of molecules of interest by linking a series of cubes, each cube having a dimension R, the locations of the cubes being determined by the calculated electron probability density of the individual one of the molecules of interest, each cube sharing at least one of its six faces with another cube, such that there is a specific number of linked cubes which varies for each individual one of the plurality of molecules of interest;

approximating the chemical reactivity of each individual one of the plurality of molecules of interest by assigning each cube of each individual one of the plurality of molecules of interest, no more than one functionality value from a plurality of M different chemical functionality values;

approximating the surface shape and chemical reactivity of a chemically active surface having a volume equal to V by subtracting a number V/R^3 cubes of dimension R from a surface, wherein each of the cube spaces shares at least one face with another cube space and wherein N of the cube spaces has one of a plurality of M different chemical functionality values;

calculating an attraction value K for each one of the plurality of molecules of interest to the chemically active surface; and

calculating a list of overall attraction values to the chemically active surface.

37. The computer of claim 36, wherein further the calculation of the attraction value K is performed on a plurality of different predetermined chemically active surfaces, and a matrix of overall attractive values of each molecule of interest to each of the different surfaces is calculated.

38. The computer of claim 36, wherein the plurality of molecules of interest includes organic molecules.

39. The computer of claim 38, wherein further the chemically active surface having a plurality of predetermined active chemical locations is calculated to correspond to the shape of an actual protein surface structure.

40. The computer of claim 36, wherein further the molecules of interest are organic molecules of 1500 Daltons or less.

41. The computer of claim 36 wherein further the chemically active surface having a plurality of predetermined active chemical locations is compared to an actual protein surface to calculate a similarity value of the actual protein surface to the predetermined active chemical locations.

42. The computer of claim 41 wherein further a plurality of predetermined chemically active surfaces are compared to a plurality of actual protein surfaces and a matrix of similarity values is calculated.

43. The computer of claim 42 wherein further the cube spaces subtracted from the surface are calculated to approximate the electron probability density of at least one of a plurality of depressions in known protein surface structures.

44. The computer of claim 42 wherein further the N sites of chemical functionality are calculated to approximate the location and type of chemical functionality of actual depressions in known protein structures.

Figure 1

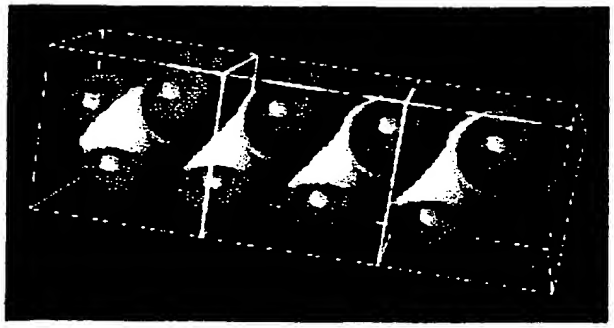


Figure 2

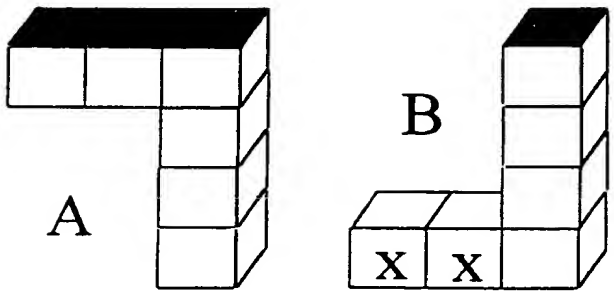


Figure 3

Figure 4

Figure 5

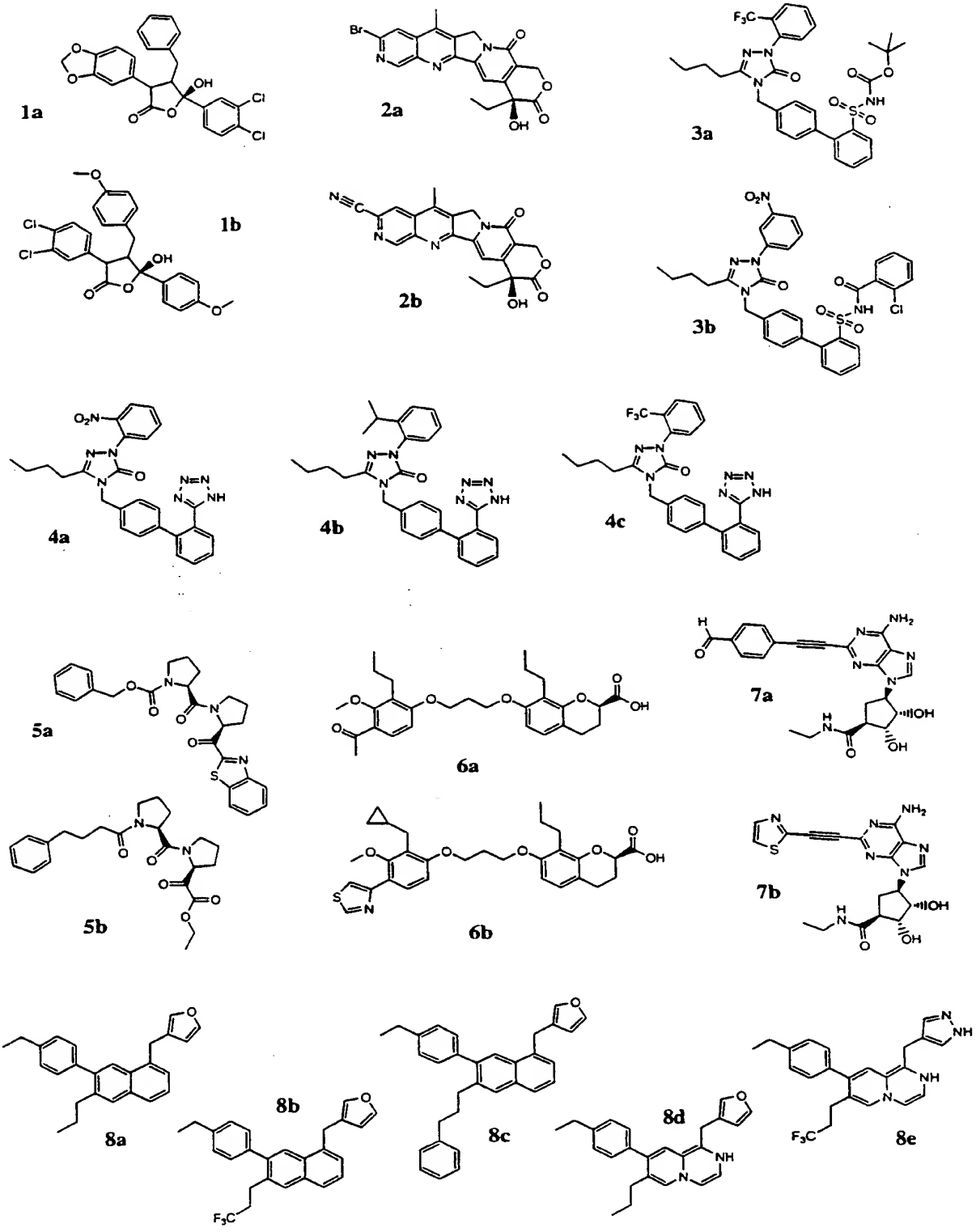


Figure 6

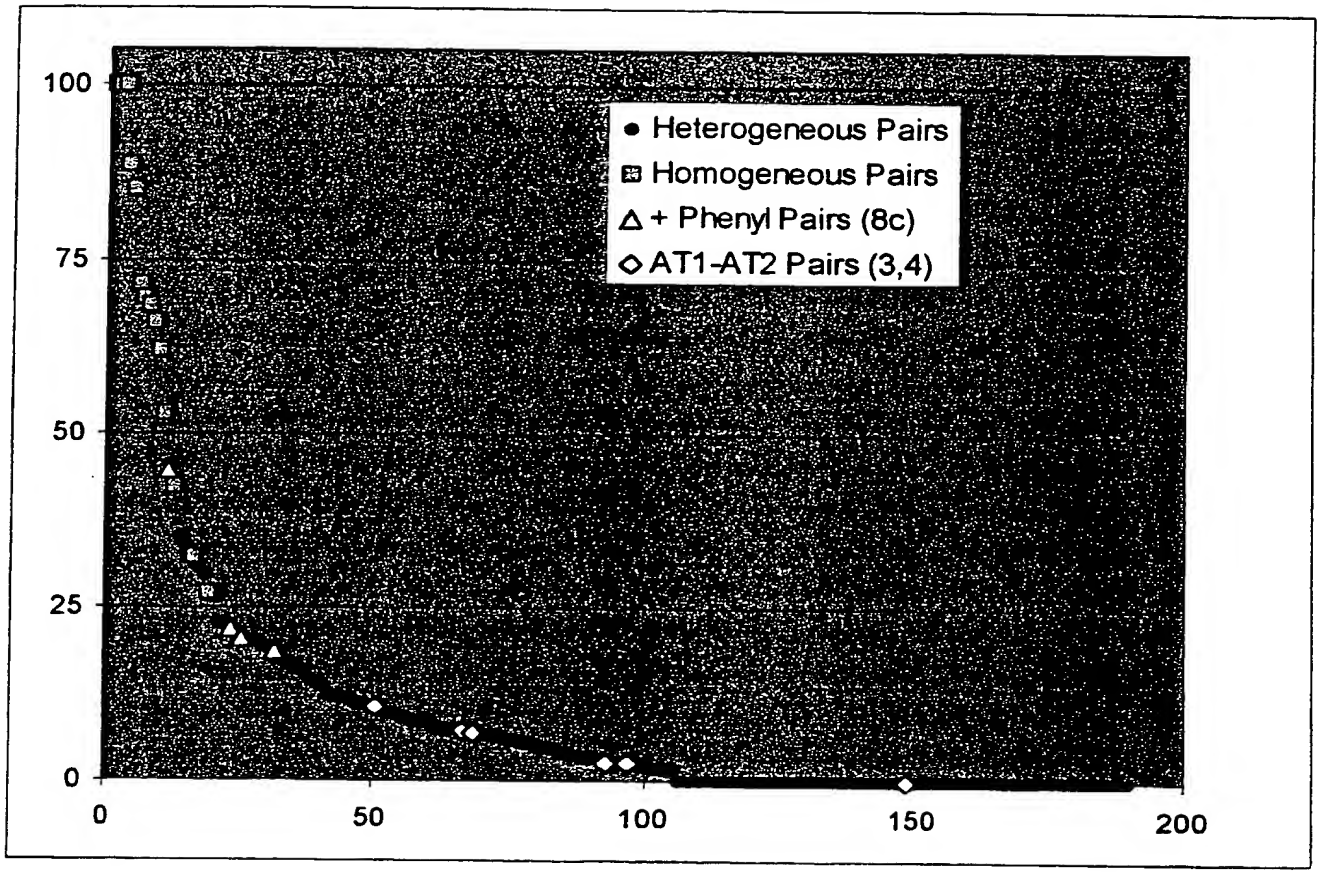
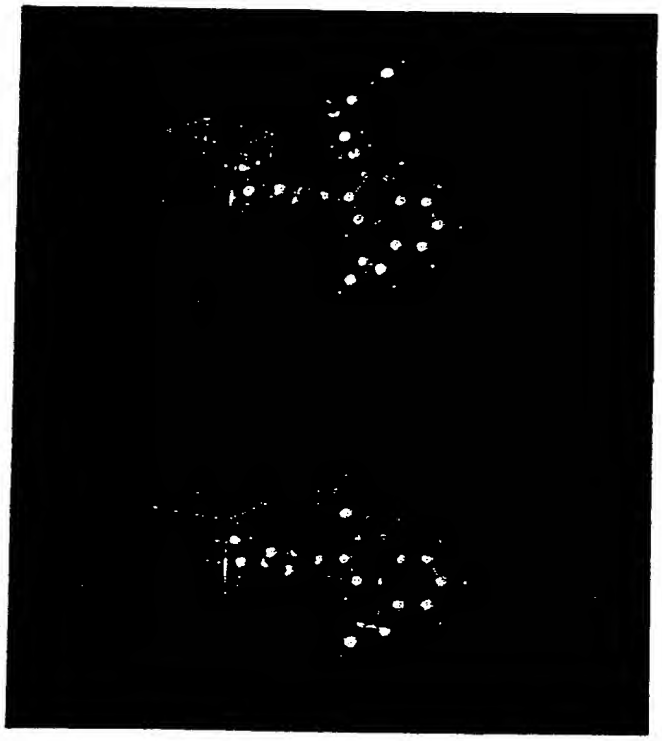

100
75
50
25
0
0
50
100
150
200
Heterogeneous Pairs
Homogeneous Pairs
+ Phenyl Pairs (8c)
AT1-AT2 Pairs (3,4)

Figure 7



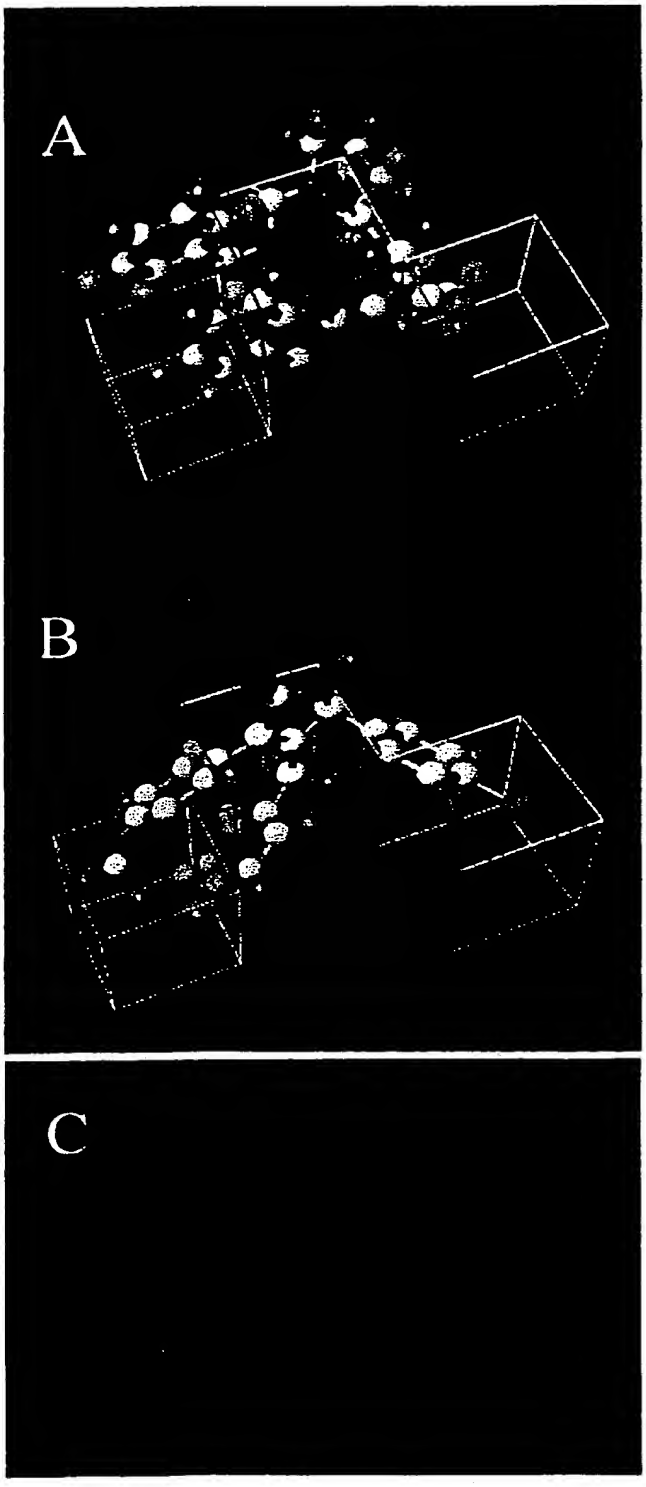


Figure 8

Figure 9

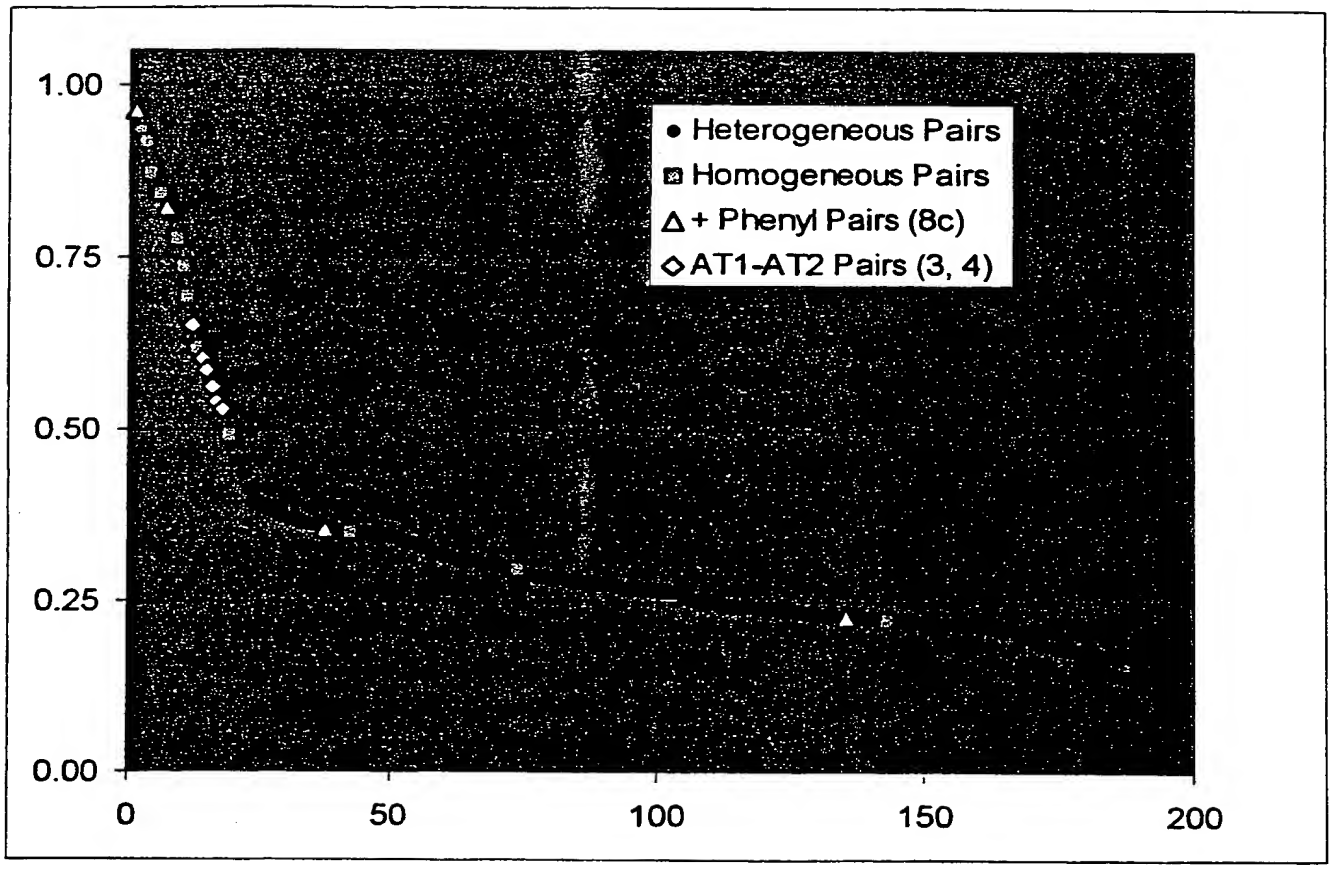


Figure 10

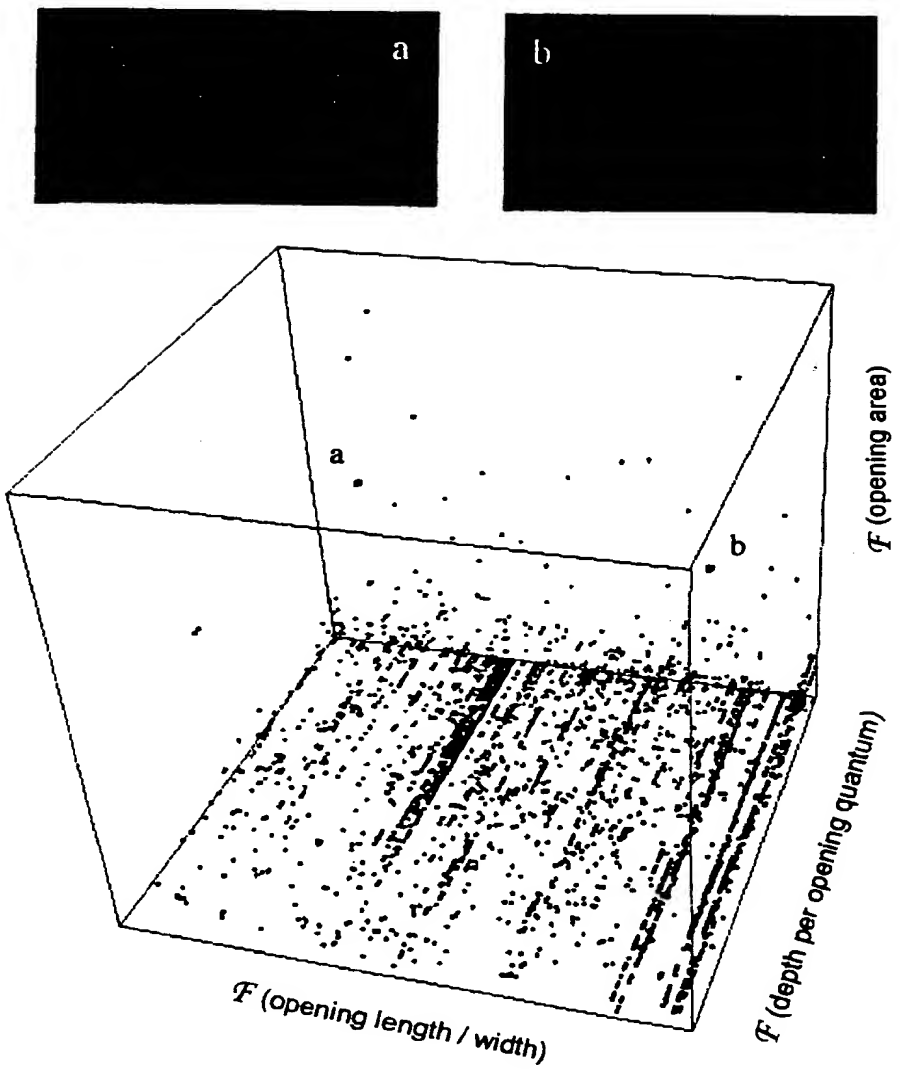


Figure 11

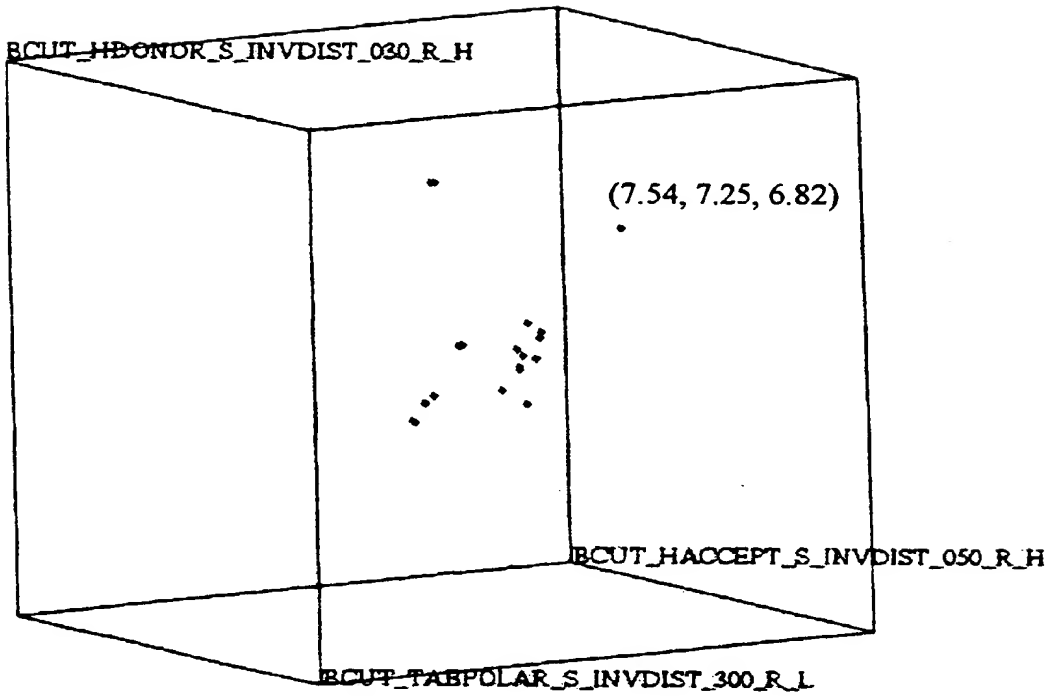
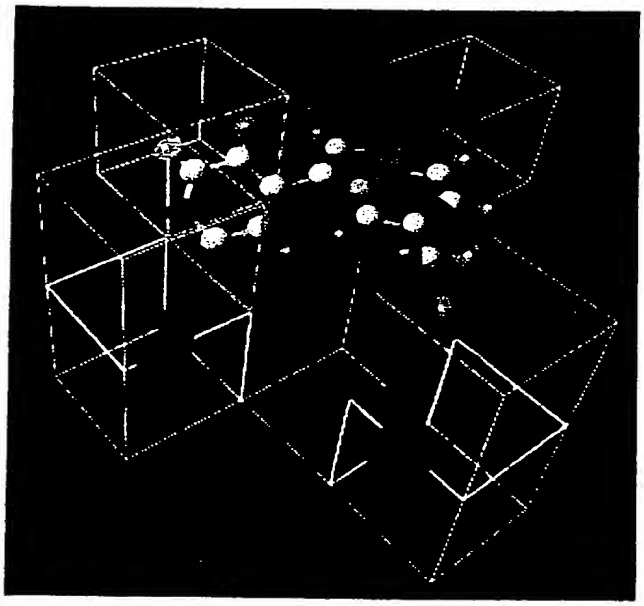


Figure 12



III

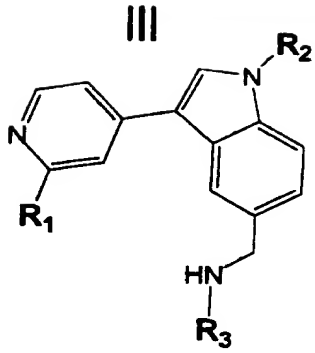
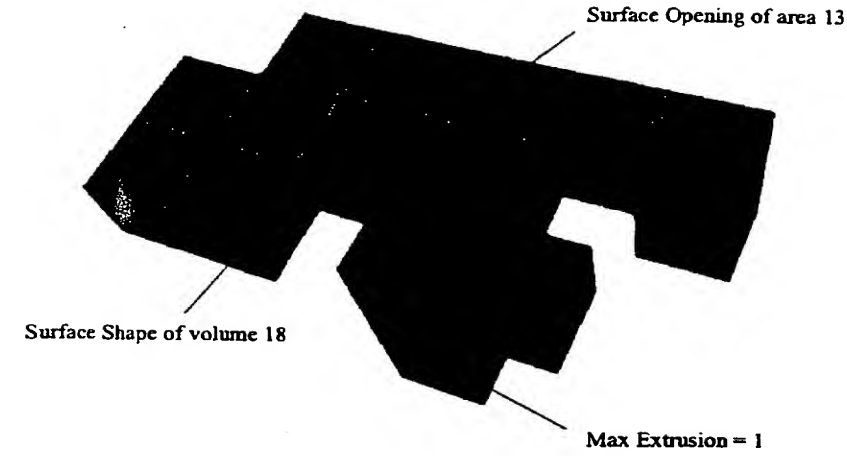


Figure 13



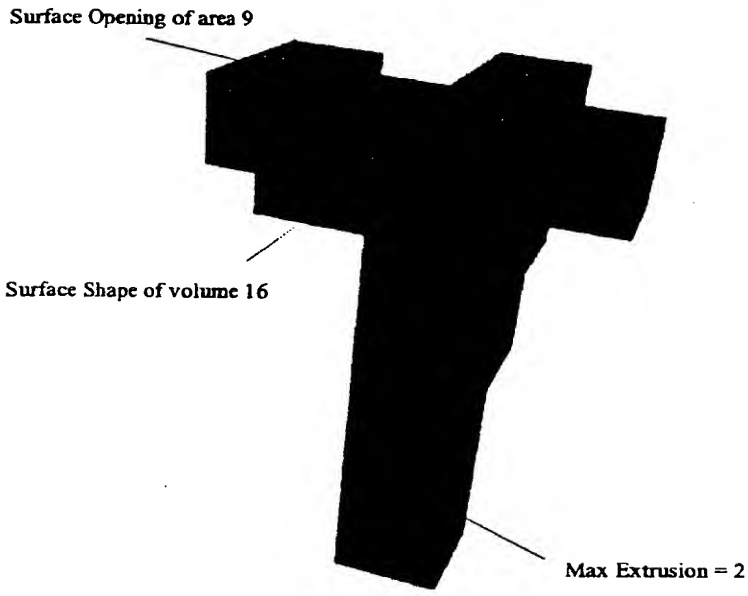


Figure 14

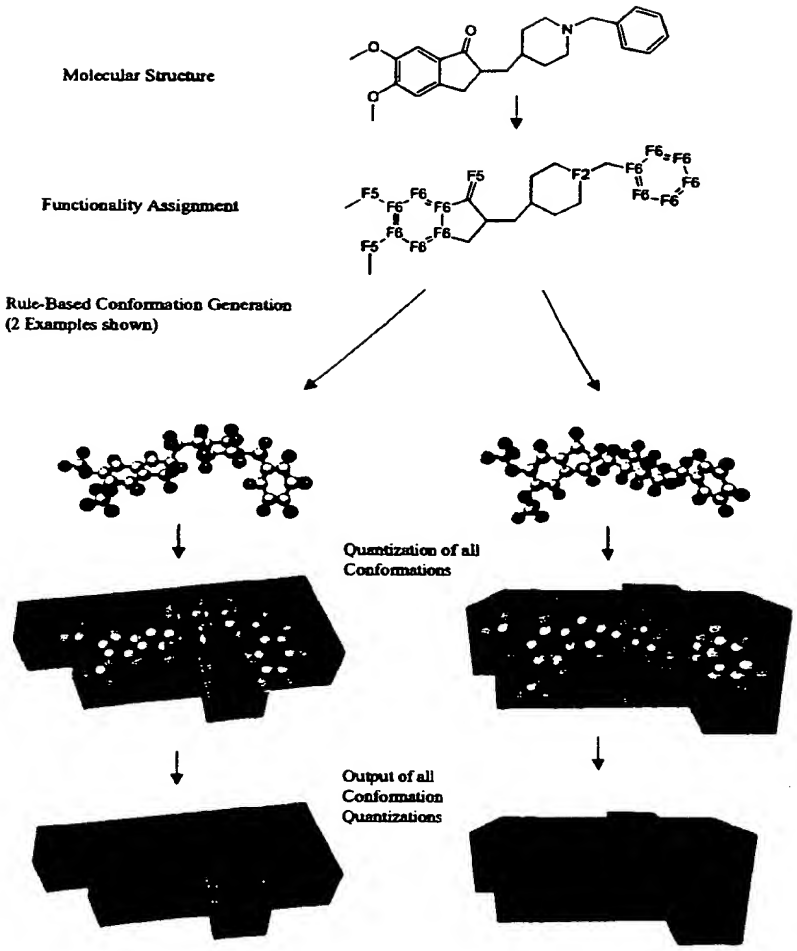


Figure 15

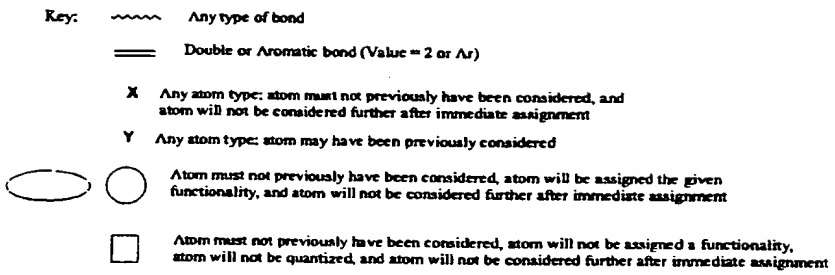


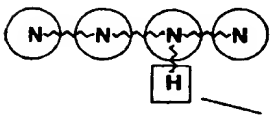
Figure 16

A ε {O.2, O.CO2}
B ε {O.3, O.CO2}

OR

A ε {O.2, O.CO2}

OR

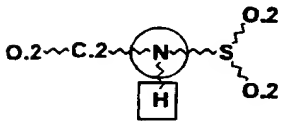


Other than bonds shown, no more than a total of two other bonds connected to all N

N ε {any type of nitrogen}

1 and only 1 H at some point on the 4-N chain

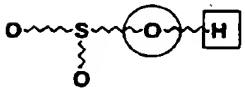
OR



N ε {any type of nitrogen}

S ε {any type of sulfur}

OR

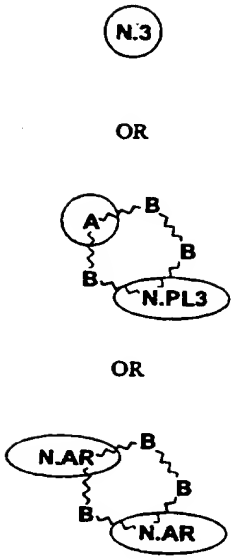


S must be bonded to exactly three oxygens

O ε {any type of oxygen}

S ε {any type of sulfur}

Figure 17



A ε {N.2, N.AR}
B ε {C.2, C.AR}

B ε {C.2, C.AR}

Figure 18

A ε {O.3, S.3}

Figure 19

A ε {N.AM,N.PL3,N.AR,N.2}
Atom must not previously have been considered, and atom will not be considered further after immediate assignment

Figure 20

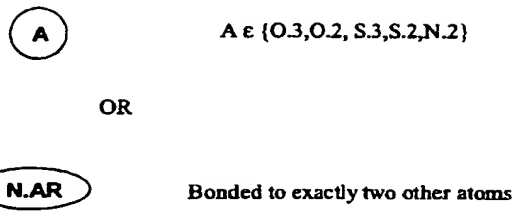


Figure 21

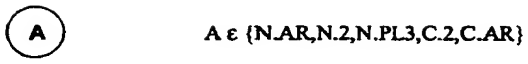


Figure 22 Table 5: Ranking of molecules in Fig. 5 by QSCD diversity score. Blue = homogeneous pairs, yellow = +phenyl pairs (8c), green = AT1-AT2 pairs (3,4)

Surface Shapes	Shape Overlaps	Shape Similarity Score	Functionality Score Per Shape Overlap	Molecule A	Molecule B	Total Similarity Score	Rank
27	27	100.00%	10	[illegible]		[illegible]	1
57	29	50.88%	10				2
107	44	41.12%	9.61				3
888	81	9.12%	9.7				4
121	11	9.09%	9.35				5
98	7	7.14%	10				6
119	9	7.56%	9.2				7
102	7	6.86%	10				8
657	54	8.22%	8.04				9
376	30	7.98%	7.76				10
716	39	5.45%	9.74				11
402	18	4.48%	10	8a	8c	44.8	12
922	39	4.23%	10				13
687	26	3.78%	9.31	1a	5a	35.2	14
570	24	4.21%	8.15	8a	5b	34.3	15
602	29	4.82%	8.77				16
933	33	3.54%	9.12				17
724	27	3.73%	8.22	3a	5a	30.7	18
727	23	3.16%	9.38	1a	5b	29.7	19
705	24	3.40%	7.94				20
535	22	4.11%	6.49	8a	5a	26.7	21
786	22	2.87%	7.94	3a	5b	22.8	22
693	18	2.31%	9.79	1a	8c	22.6	23
366	8	2.19%	10	8b	8c	21.9	24
607	16	3.16%	6.79	8b	5a	21.4	25
389	8	2.06%	10	8d	8c	20.6	26
675	18	2.37%	8.55	3b	1a	20.3	27
289	7	2.42%	8.3	7b	2a	20.1	28
289	7	2.42%	8.3	7b	2b	20.1	29
519	18	3.08%	6.3	3b	8a	19.4	30
548	14	2.65%	7.54	1a	8b	19.3	31
371	7	1.89%	10	8a	8c	18.9	32
801	15	1.87%	9.28	4b	5a	17.4	33
547	13	2.38%	7.27	8b	5b	17.3	34
776	14	1.80%	9.5	3a	1a	17.1	35
618	16	2.59%	6.3	3a	8a	16.3	36
809	15	1.85%	8.74	4a	5a	16.2	37
568	12	2.04%	7.47	3a	8b	15.2	38
636	16	2.52%	5.73	3b	5a	14.4	39
741	12	1.62%	8.74	1b	5b	14.1	40
843	12	1.42%	9.71	4b	1a	13.8	41
383	6	1.57%	7.94	7a	2a	12.4	42
383	6	1.57%	7.94	7a	2b	12.4	43
807	12	1.49%	8.36	4c	5a	12.4	44
847	11	1.30%	9.35	4c	1a	12.1	45
454	9	1.98%	6.06	1a	8a	12.0	46
846	10	1.18%	9.86	4b	1b	11.4	47
739	17	2.30%	4.9	7b	4a	11.3	48
704	8	1.14%	9.57	1b	8c	10.9	49
842	11	1.31%	8.17	4b	5b	10.7	50
882	11	1.25%	8.6	4b	3a	10.7	51
829	15	1.81%	5.85	7a	4c	10.6	52
700	7	1.00%	10	8c	5b	10.0	53
853	10	1.17%	8.44	4a	1a	9.9	54
740	11	1.48%	6.49	7b	4c	9.6	55
589	7	1.19%	7.54	1a	8a	9.0	56
663	7	1.06%	8.3	8c	5a	8.8	57
417	4	0.96%	9.09	1a	8a	8.7	58
435	5	1.15%	7.37	1a	8d	8.5	59
853	8	0.94%	9.09	4c	1b	8.5	60
709	7	0.99%	8.3	1b	5a	8.2	61
833	16	1.92%	3.97	7a	4a	7.6	62
398	3	0.75%	10	2a	5b	7.5	63
398	3	0.75%	10	2b	5b	7.5	64
806	6	0.74%	10	4b	8c	7.4	65
809	6	0.74%	10	4c	8c	7.4	66
794	8	1.01%	7.21	4a	3b	7.3	67
854	7	0.82%	8.94	4a	5b	7.3	68
893	8	0.90%	7.94	4a	3a	7.1	69
883	10	1.51%	4.65	4a	8b	7.0	70
398	3	0.75%	8.74	8d	5a	6.6	71
880	6	0.70%	9.41	4a	1b	6.6	72
680	9	1.32%	4.81	3b	5b	6.4	73
785	8	1.02%	6.3	3a	1b	6.4	74
414	3	0.72%	8.74	1a	8b	6.3	75
436	3	0.69%	8.74	8d	5b	6.0	76
621	3	0.58%	10	4b	8a	5.8	77
681	4	0.71%	7.94	1b	8b	5.7	78
483	3	0.65%	8.74	1b	8a	5.7	79
834	7	0.84%	6.59	7a	4b	5.5	80
642	6	0.93%	5.51	7b	1b	5.1	81
850	6	0.71%	6.94	4c	5b	4.9	82
644	4	0.62%	7.94	3b	8c	4.9	83
422	2	0.47%	10	1b	8a	4.7	84
512	2	0.39%	10	4a	2a	3.9	85
512	2	0.39%	10	4a	2b	3.9	86
422	2	0.47%	7.94	8a	5a	3.8	87
541	2	0.37%	10	4b	8d	3.7	88
695	4	0.58%	6.3	7a	5a	3.6	89
743	4	0.54%	6.3	3a	8c	3.4	90
597	2	0.34%	10	1b	8a	3.4	91
817	3	0.37%	8.74	4a	8c	3.2	92
892	4	0.45%	6.3	4c	3a	2.8	93
359	1	0.28%	10	2a	8c	2.8	94
359	1	0.28%	10	2b	8c	2.8	95
377	1	0.27%	10	8b	5a	2.7	96
791	3	0.38%	6.94	4b	3b	2.6	97
381	1	0.26%	10	8a	5a	2.6	98
866	2	0.30%	7.94	4c	8b	2.4	99
442	1	0.23%	10	1b	8d	2.3	100
508	1	0.20%	10	4c	2a	2.0	101
508	1	0.20%	10	4c	2b	2.0	102
519	1	0.19%	10	4b	8b	1.9	103
565	1	0.18%	10	4b	8a	1.8	104
552	1	0.18%	10	8a	8c	1.8	105
633	12	1.90%	0	7b	1a	0.0	106
492	9	1.83%	0	3b	8b	0.0	107
210	3	1.43%	0	2a	8b	0.0	108
210	3	1.43%	0	2b	8b	0.0	109
449	6	1.34%	0	7b	8b	0.0	110
599	7	1.17%	0	7b	8a	0.0	111
741	7	0.94%	0	7b	4b	0.0	112
543	5	0.92%	0	7a	8b	0.0	113
696	6	0.86%	0	4c	8a	0.0	114
638	5	0.78%	0	7b	5b	0.0	115
678	5	0.74%	0	7b	3a	0.0	116
702	5	0.71%	0	4a	8a	0.0	117
578	4	0.69%	0	7a	8a	0.0	118
771	5	0.65%	0	7a	3a	0.0	119
488	3	0.62%	0	7b	8a	0.0	120
690	4	0.58%	0	3b	1b	0.0	121
354	2	0.56%	0	7b	8c	0.0	122
581	3	0.52%	0	7b	3b	0.0	123
226	1	0.44%	0	8b	8b	0.0	124
459	2	0.44%	0	8a	5b	0.0	125
696	3	0.43%	0	4b	8a	0.0	126
230	1	0.43%	0	8b	8a	0.0	127
738	3	0.41%	0	7a	1b	0.0	128
517	2	0.39%	0	8b	8c	0.0	129
283	1	0.35%	0	8a	8d	0.0	130
306	1	0.33%	0	8a	8a	0.0	131
313	1	0.32%	0	7b	8a	0.0	132
675	2	0.30%	0	7a	3b	0.0	133
332	1	0.30%	0	7b	8d	0.0	134
663	2	0.30%	0	4b	8b	0.0	135
734	2	0.27%	0	7a	5b	0.0	136
405	1	0.25%	0	2a	1b	0.0	137
402	1	0.25%	0	2a	1a	0.0	138
405	1	0.25%	0	2b	1b	0.0	139
402	1	0.25%	0	2b	1a	0.0	140
425	1	0.24%	0	7a	8d	0.0	141
418	1	0.24%	0	8a	5b	0.0	142
414	1	0.24%	0	8b	5b	0.0	143
448	1	0.22%	0	7a	8a	0.0	144
505	1	0.20%	0	4b	2a	0.0	145
505	1	0.20%	0	4b	2b	0.0	146
601	1	0.17%	0	7b	8c	0.0	147
737	1	0.14%	0	7a	1a	0.0	148
786	1	0.13%	0	4c	3b	0.0	149
407	0	0.00%	0	7a	8a	0.0	150
403	0	0.00%	0	7a	8b	0.0	151
695	0	0.00%	0	7a	8c	0.0	152
310	0	0.00%	0	7b	8b	0.0	153
527	0	0.00%	0	4c	8a	0.0	154
546	0	0.00%	0	4c	8d	0.0	155
569	0	0.00%	0	4c	8a	0.0	156
523	0	0.00%	0	4c	8b	0.0	157
632	0	0.00%	0	4a	8a	0.0	158
551	0	0.00%	0	4a	8d	0.0	159
574	0	0.00%	0	4a	8a	0.0	160
528	0	0.00%	0	4a	8b	0.0	161
342	0	0.00%	0	3b	2a	0.0	162
342	0	0.00%	0	3b	2b	0.0	163
360	0	0.00%	0	3b	8a	0.0	164
379	0	0.00%	0	3b	8d	0.0	165
402	0	0.00%	0	3b	8a	0.0	166
356	0	0.00%	0	3b	8b	0.0	167
441	0	0.00%	0	3a	2a	0.0	168
441	0	0.00%	0	3a	2b	0.0	169
459	0	0.00%	0	3a	8a	0.0	170
478	0	0.00%	0	3a	8d	0.0	171
501	0	0.00%	0	3a	8a	0.0	172
455	0	0.00%	0	3a	8b	0.0	173
247	0	0.00%	0	2a	8a	0.0	174
72	0	0.00%	0	2a	8a	0.0	175
91	0	0.00%	0	2a	8d	0.0	176
114	0	0.00%	0	2a	8a	0.0	177
68	0	0.00%	0	2a	8b	0.0	178
364	0	0.00%	0	2a	5a	0.0	179
247	0	0.00%	0	2b	8a	0.0	180
72	0	0.00%	0	2b	8a	0.0	181
91	0	0.00%	0	2b	8d	0.0	182
114	0	0.00%	0	2b	8a	0.0	183
68	0	0.00%	0	2b	8b	0.0	184
364	0	0.00%	0	2b	5a	0.0	185
420	0	0.00%	0	1b	8b	0.0	186
265	0	0.00%	0	8a	8a	0.0	187
261	0	0.00%	0	8a	8b	0.0	188
250	0	0.00%	0	8b	8d	0.0	189
273	0	0.00%	0	8b	8a	0.0	190

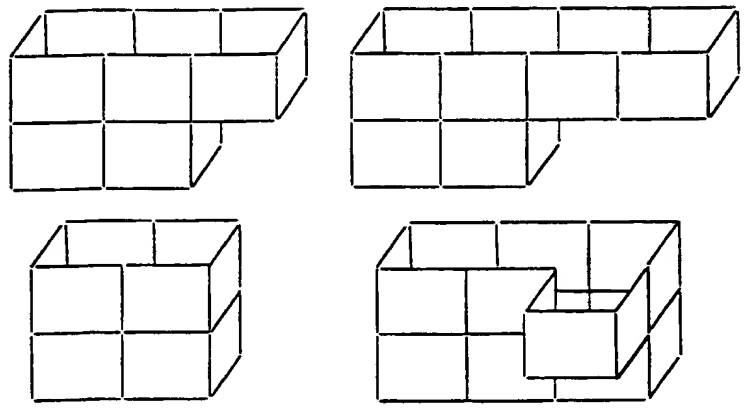
Figure 23 Table 6: Ranking of molecules in Fig. 5 by Tanimoto similarity score of 2D UNITY fingerprints. Blue = homogeneous pairs, yellow = +phenyl pairs (8c), green = AT1-AT2 pairs (3,4)

AND	OR	Molecule A	Molecule B	Tanimoto Score	Rank
112	116	8c	8a	0.97	1
370	397	3b	7a	0.93	2
195	212	4c	8c	0.92	3
276	316	4c	8a	0.87	4
112	132	8c	8b	0.85	5
275	326	7a	8a	0.84	6
112	136	8c	8b	0.82	7
268	328	4a	8a	0.82	8
265	341	4b	8c	0.78	9
186	252	4a	8c	0.74	10
292	421	7a	8c	0.66	11
260	398	3a	4c	0.65	12
152	245	3b	4c	0.65	13
249	412	3b	4a	0.60	14
238	408	3a	4b	0.59	15
233	414	3b	4b	0.56	16
232	429	3b	4c	0.54	17
228	430	3a	4a	0.53	18
157	319	8a	8c	0.48	19
118	245	6a	1b	0.48	20
118	260	6a	1a	0.45	21
129	297	6b	1b	0.43	22
129	312	6b	1a	0.41	23
192	486	2b	7a	0.40	24
101	257	5b	6a	0.39	25
191	494	2a	7a	0.39	26
188	495	2b	7b	0.38	27
184	488	5a	2a	0.38	28
181	484	5a	2b	0.37	29
159	429	5a	7b	0.37	30
186	504	2a	7b	0.37	31
108	294	5b	1a	0.37	32
156	427	5a	7a	0.37	33
103	284	5b	1b	0.36	34
157	434	8d	2b	0.36	35
177	492	2b	4b	0.36	36
86	241	8c	8d	0.36	37
157	441	8d	2a	0.36	38
156	449	8a	2b	0.35	39
134	381	8a	8b	0.35	40
84	239	8a	8c	0.35	41
89	254	8b	8d	0.35	42
159	455	8a	2a	0.35	43
176	507	2b	4a	0.35	44
126	383	5b	7a	0.35	45
191	551	2a	3a	0.35	46
174	502	2a	4b	0.35	47
189	546	2b	3a	0.35	48
188	488	3b	7a	0.34	49
189	549	2b	3b	0.34	50
153	448	4a	7a	0.34	51
189	556	2a	3b	0.34	52
108	315	5b	8b	0.34	53
173	517	2a	4a	0.33	54
147	440	4b	7a	0.33	55
170	513	2b	4c	0.33	56
132	401	6b	7b	0.33	57
168	522	2a	4c	0.32	58
146	460	4a	7b	0.32	59
139	439	5b	2a	0.32	60
146	464	6b	2b	0.31	61
153	487	5a	3a	0.31	62
158	503	3b	7b	0.31	63
143	458	4c	7a	0.31	64
140	452	4b	7b	0.31	65
135	436	5b	2b	0.31	66
153	500	3a	7a	0.31	67
144	473	6b	2a	0.30	68
137	451	5a	4a	0.30	69
115	379	5b	7b	0.30	70
123	407	8a	4c	0.30	71
122	406	6b	7a	0.30	72
147	486	5a	3b	0.30	73
82	277	8b	8c	0.29	74
112	384	5a	1a	0.29	75
136	470	4c	7b	0.29	76
128	446	5a	4b	0.29	77
115	401	8a	4b	0.29	78
128	448	1b	2b	0.29	79
146	512	3a	7b	0.29	80
131	460	1a	2b	0.28	81
116	404	6b	4b	0.28	82
106	375	5a	1b	0.28	83
106	386	5b	4a	0.28	84
111	396	1a	7a	0.28	85
127	456	1b	2a	0.28	86
116	417	6b	4a	0.28	87
130	468	1a	2a	0.28	88
127	458	6b	3a	0.28	89
107	387	1b	7a	0.28	90
125	457	8a	3a	0.27	91
126	462	5a	4c	0.27	92
107	393	8d	4b	0.27	93
102	378	5b	4b	0.27	94
116	430	5b	3a	0.27	95
124	461	8a	3b	0.27	96
112	416	8a	4a	0.27	97
98	367	6a	7a	0.27	98
95	357	5a	6a	0.27	99
112	421	6b	4c	0.27	100

AND	OR	Molecule A	Molecule B	Tanimoto Score	Rank
115	434	5b	3b	0.26	101
119	450	8d	3b	0.26	102
84	318	6b	8d	0.26	103
114	433	8a	2b	0.26	104
122	466	6b	3b	0.26	105
104	410	8d	4a	0.25	106
104	410	8d	4c	0.25	107
112	442	8a	2a	0.25	108
83	335	5b	8a	0.25	109
98	386	5b	4c	0.25	110
104	421	8a	7a	0.25	111
58	227	8c	6a	0.25	112
111	455	8d	3a	0.24	113
97	399	5a	8d	0.24	114
61	251	8c	1b	0.24	115
85	350	8b	4c	0.24	116
100	412	5a	8a	0.24	117
79	328	8c	4b	0.24	118
54	225	8a	6a	0.24	119
110	456	1a	3b	0.24	120
99	415	1a	7b	0.24	121
109	457	1a	3a	0.24	122
63	265	8b	1b	0.24	123
77	324	8a	4b	0.24	124
59	249	8a	1b	0.24	125
97	412	8d	7a	0.24	126
95	404	1b	7b	0.24	127
85	405	1a	4b	0.23	128
89	381	6a	4a	0.23	129
86	370	6a	4b	0.23	130
104	450	1b	3b	0.23	131
78	344	8a	1b	0.23	132
76	331	8d	1b	0.23	133
98	427	8a	3b	0.23	134
84	279	8c	8a	0.23	135
61	268	8c	1a	0.23	136
78	343	8b	4b	0.23	137
102	449	1b	3a	0.23	138
64	282	8c	6b	0.23	139
55	244	8b	8a	0.23	140
81	360	8d	6b	0.23	141
63	280	8b	1a	0.23	142
62	277	8b	8c	0.23	143
94	420	1a	4a	0.22	144
59	264	8a	1a	0.22	145
85	427	6a	3a	0.22	146
88	397	1b	4b	0.22	147
78	343	8c	4a	0.22	148
62	280	8a	6b	0.22	149
79	359	8a	1a	0.22	150
78	348	8d	1a	0.22	151
82	375	8a	6b	0.22	152
55	252	5b	8c	0.22	153
84	386	6a	7b	0.22	154
74	341	8a	4a	0.22	155
94	436	8a	7b	0.22	156
67	311	8d	6a	0.22	157
64	298	8b	6b	0.21	158
74	345	8c	4c	0.21	159
83	387	6a	4c	0.21	160
88	411	1b	4a	0.21	161
90	424	1a	4c	0.21	162
53	250	5b	8a	0.21	163
72	343	8a	4c	0.21	164
84	403	8b	3a	0.21	165
75	380	8b	4a	0.21	166
87	427	8d	7b	0.20	167
66	328	8a	6a	0.20	168
83	416	1b	4c	0.20	169
53	270	5b	8b	0.20	170
75	399	8c	3b	0.19	171
77	413	8b	3b	0.19	172
77	419	8c	2b	0.18	173
79	433	8b	2b	0.18	174
72	398	8a	3b	0.18	175
75	417	8a	2b	0.18	176
63	351	8c	7a	0.18	177
78	441	8b	2a	0.18	178
60	341	5a	8c	0.18	179
75	428	8c	2a	0.18	180
61	349	8a	7a	0.17	181
59	338	5a	8a	0.17	182
63	367	8b	7a	0.17	183
69	402	8c	3a	0.17	184
73	426	8a	2a	0.17	185
61	356	6a	8b	0.17	186
68	399	8a	3a	0.17	187
52	367	8c	7b	0.14	188
50	365	8a	7b	0.14	189
52	383	8b	7b	0.14	190

Figure 24

Example subset of theoretical surfaces Ti containing 4 members:



Example Central set Ci for Ti (F = 1, E = 3)
where black face denotes a point of attachment A1 on Ci:

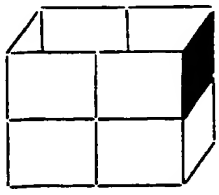


Figure 25

Example Core Molecule Mi to fill Central set Ci:

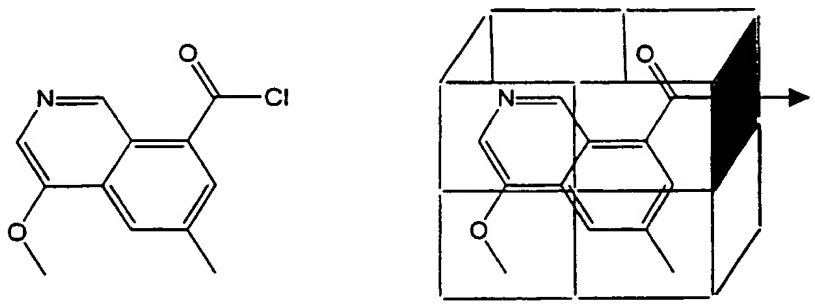


Figure 26

Example Library L(Mi,B) where B = a set of amines:

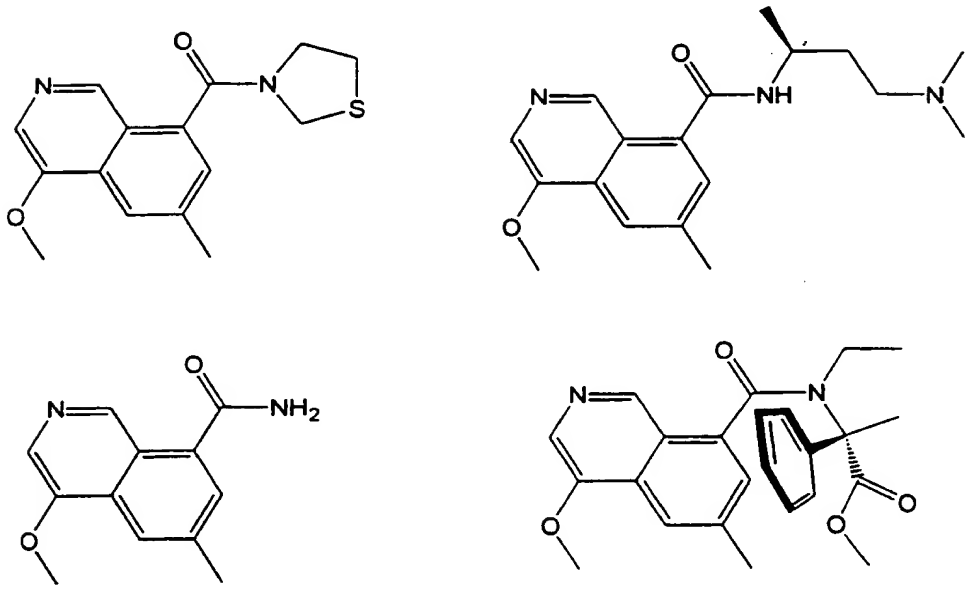
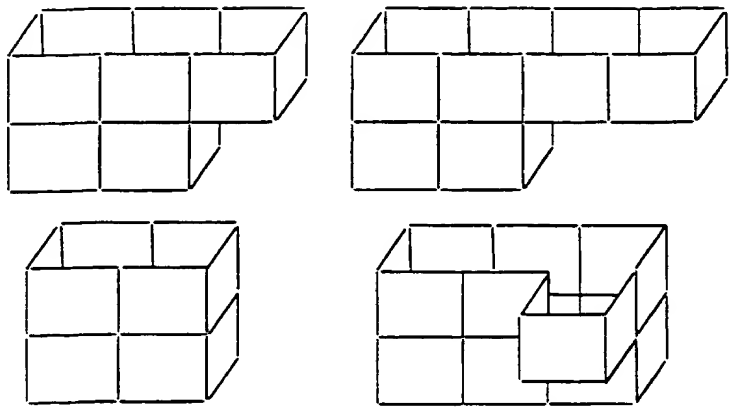


Figure 27

Example subset of target surfaces Ti containing 4 members:



Example Central set Ci for Ti (F = 1, E = 3)
where black face denotes a point of attachment A1 on Ci:

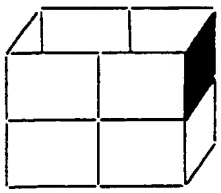


Figure 28

Example Core Molecule Mi to fill Central set Ci:

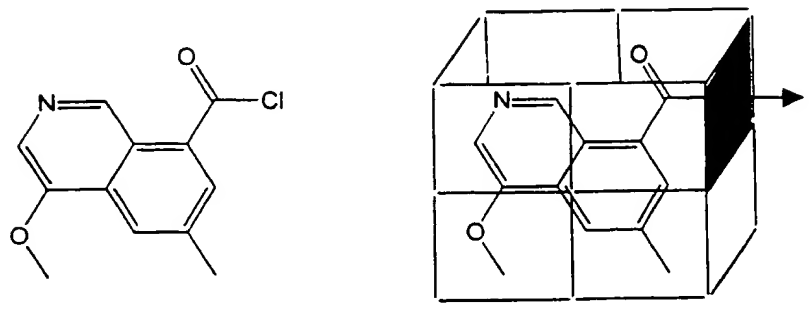


Figure 29

Example Library L(Mi,B) where B = a set of amines:

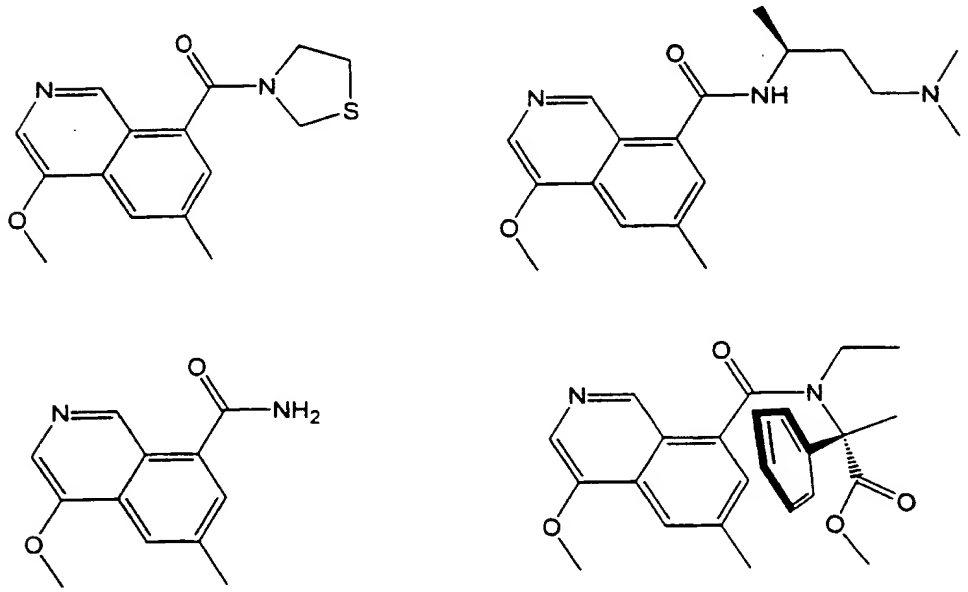
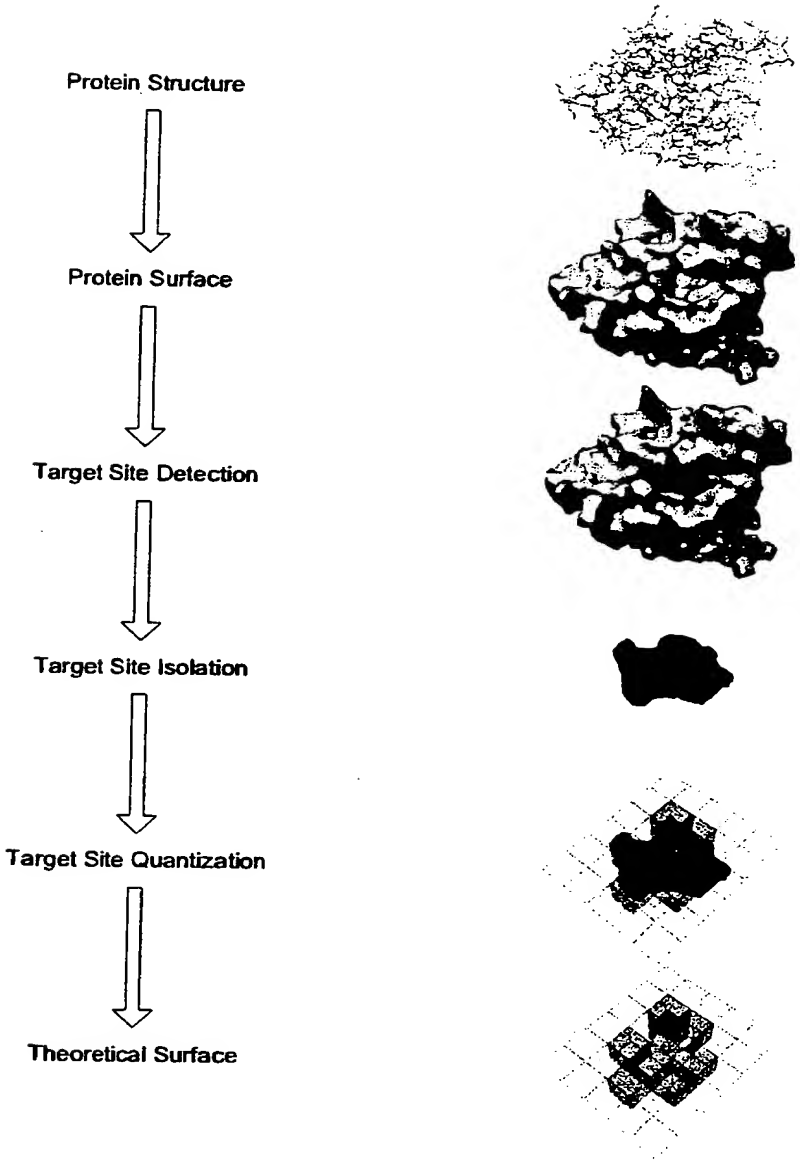


Figure 30



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 October 2000 (12.10.2000)

PCT

(10) International Publication Number
WO 00/60507 A3

(51) International Patent Classification[7]: G06F 17/50

(21) International Application Number: PCT/US00/08777

(22) International Filing Date: 31 March 2000 (31.03.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/127,486 2 April 1999 (02.04.1999) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/127,486 (CIP)
Filed on 2 April 1999 (02.04.1999)

(71) Applicant *(for all designated States except US)*: NEOGENESIS, INC. [US/US]; 840 Memorial Drive, Cambridge, MA 02139 (US).

(72) Inventors; and

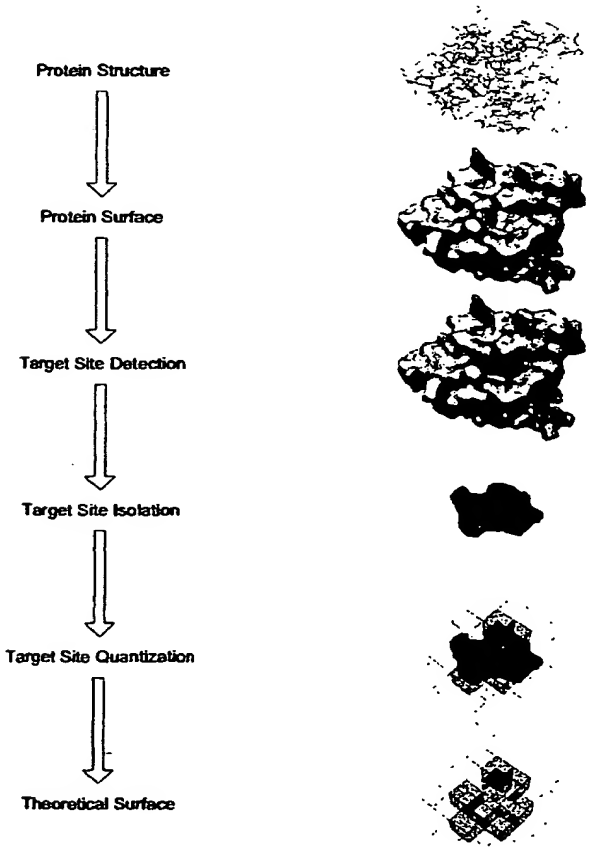
(75) Inventors/Applicants *(for US only)*: WINTNER, Edward, A. [US/US]; 44 Valentine Street, Cambridge, MA 02139 (US). MOALLEMI, Ciamac, C. [US/US]; Apartment 2-4A, 100 Memorial Drive, Cambridge, MA 02142 (US).

(74) Agent: KLUNDER, Janice, M.; Hale and Dorr, LLP, 60 State Street, Boston, MA 02109 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ,

[Continued on next page]

(54) Title: ANALYZING MOLECULE AND PROTEIN DIVERSITY



(57) Abstract: A computer-based method in which a set of constraints is placed on possible target surfaces, and a fully enumerated set of theoretical target surfaces under the given constraints is created, such that each surface has a defined, continuous volume and a defined, continuous surface area. One or more sets of objects are mapped to the fully enumerated set of theoretical target surfaces to define corresponding subsets of the fully enumerated set of theoretical target surfaces. An aspect of diversity of the objects is analyzed based on degrees of similarities and differences among the corresponding subsets.

WO 00/60507 A3

PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *With international search report.*

(88) Date of publication of the international search report:
12 April 2001

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

Intern ıal Application No
PCT/US 00/08777

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/50

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
BIOSIS, EPO-Internal, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JOHN MOUNT ET AL: "Icepick: A flexible surface-based system for molecular diversity" JOURNAL OF MEDICINAL CHEMISTRY, 'Online! vol. 42, no. 1, 1999, pages 60-66, XP002156348 Retrieved from the Internet: <URL:http://pubs3.acs.org/s97is.vts> 'retrieved on 2000-12-28! published on the Internet on the 12/24/1998 page 60 -page 62, column 2, line 17 --- -/--	1,34-36

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

'A' document defining the general state of the art which is not considered to be of particular relevance

'E' earlier document but published on or after the international filing date

'L' document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

'O' document referring to an oral disclosure, use, exhibition or other means

'P' document published prior to the international filing date but later than the priority date claimed

'T' later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

'X' document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

'Y' document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

'&' document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 December 2000	11/01/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Guingale, A

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Interr nal Application No
PCT/US 00/08777

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	PARKS CAMDEN A ET AL: "The measurement of molecular diversity by receptor site interaction simulation." JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN, vol. 12, no. 5, 1998, pages 441-449, XP000972384 ISSN: 0920-654X page 441 -page 449, column 1, line 29 ---	1,34-36
A	BURKHARD P ET AL: "An example of a protein ligand found by database mining: Description of the docking method and its verification by a 2.3 A X-ray structure of a thrombin-ligand complex." JOURNAL OF MOLECULAR BIOLOGY, vol. 277, no. 2, 27 March 1998 (1998-03-27), pages 449-466, XP000974379 ISSN: 0022-2836 page 449 -page 454, column 1, line 6 ---	1,34-36
A	JIANG F ET AL: "SOFT DOCKING MATCHING OF MOLECULAR SURFACE CUBES" JOURNAL OF MOLECULAR BIOLOGY, vol. 219, no. 1, 1991, pages 79-102, XP000972336 ISSN: 0022-2836 page 79 -----	1,34-36

Form PCT/ISA/210 (continuation of second sheet) (July 1992)